The Infection Dynamics of Insertion Sequences

Dissertation

zur
Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der
Mathematisch-naturwissenschaftlichen Fakultät
der
Universität Zürich

von
Manuel Bichsel
von
Sumiswald BE

Promotionskomitee:
Prof. Dr. Andreas Wagner (Vorsitz)
Prof. Dr. Andrew D. Barbour
Prof. Dr. Hanna Kokko

Zürich, 2015

Contents

Summary

Mobile DNA encompasses all genetic elements that can move within or between genomes of their host organisms. In general, those elements persist not because they fulfill an adaptive function for their host, but because they can increase their number in a genome and infect new genomes. Mobile DNA can thus basically be regarded as a genomic parasite, although it may be co-opted by its host to fulfill a function.

Bacterial insertion sequences (ISs) are the simplest form of autonomous mobile DNA, in that they usually contain only one gene, which encodes the enzyme needed for their mobility (transposition) inside their prokaryote host genome. Through their transposition activity and through ectopic recombination, which can occur if more than one IS copy is present in a host genome, ISs have in general a detrimental fitness effect on their prokaryote hosts and depend on horizontal gene transfer (HGT) to persist by infecting new hosts.

In my thesis, I analysed the infection dynamics of an IS that has been introduced into an uninfected host cell population. First, I modeled the early phase of such an IS infection that starts with an uninfected host cell population at carrying capacity, containing just a single cell infected with a slightly detrimental IS. I found that the infection process is very erratic in its early phase, that the IS infection has a very high probability of becoming extinct and that in fact most new IS infections become extinct during the very first few cell generations. My analysis showed that for an IS infection that survives the early phase, the prevalence of infected cells in the host population grows only slowly. According to my calculations, most infected cells contain only few IS copies, which is in good agreement with the IS count distribution that I found in 728 fully sequenced prokaryotic genomes. In a second analysis, I turned to the late phase of an IS infection that has survived the early phase. I modeled the equilibrium distribution of IS counts per genome in the host cell population. Based on this model and on the IS count distribution of a specific IS familiy in 1128 fully sequenced prokaryote genomes, I calculated maximum likelihood estimates for the HGT rate and for the fitness effect of an IS. I found that the ISs in this family are at most slightly detrimental and might even be effectively neutral, and that the estimated HGT rate is in the range of HGT rates observed in the wild. Using the estimated parameters to determine

the IS infection dynamics, I found that the time needed to reach the prevalence of infected cells that can be observed in the wild is unrealistically long. Therefore I conjectured that beneficial fitness effects (if only occasional or temporary) may have played an important role in the quick spreading of ISs. Finally, in a third project, I simulated the spreading of an IS that can have both detrimental and beneficial fitness effects on its host cells in metapopulations consisting of spatially distributed subpopulations. I observed that the (beneficial) fitness effect of an IS and the HGT rate in a subpopulation strongly influence the speed with which the IS infection spreads in a metapopulation, but that the spatial structure of the metapopulation and dispersal between subpopulations do not seem to strongly limit the speed of IS spreading. My simulations showed that the initially infected subpopulation plays a crucial role in the infection dynamics of a metapopulation.

Zusammenfassung

Mobile DNA umfasst sämtliche genetischen Elemente, welche sich innerhalb eines Genoms oder zwischen den Genomen von Wirtsorganismen bewegen können. Diese Elemente bleiben im Wirtsgenom nicht bestehen weil sie eine adaptive Funktion für den Wirtsorganismus erfüllen, sondern weil sie sich innerhalb eines Genoms vermehren und neue Genome infizieren können. Mobile DNA kann deshalb als Parasit des Genoms betrachtet werden.

Bakterielle Insertionssequenzen (IS) stellen die einfachste Form von autonomer, mobiler DNA dar. Sie enthalten üblicherweise nur ein Gen, welches ein Enzym codiert, das wiederum der Mobilität im Wirtsgenom dient (Transposition). Wegen ihrer Mobilität und wegen der Möglichkeit zur ektopischen Rekombination bei mehreren IS im Genom reduzieren IS im Allgemeinen die Fitness ihres Wirtsorganismus. Sie benötigen deshalb horizontalen Gentransfer (HGT), der ihren Fortbestand durch die Infektion neuer Wirtsorganismen sichert.

In meiner Dissertation habe ich die Infektionsdynamik einer neu in eine Wirtspopulation eingeführten IS untersucht. Zuerst habe ich die Frühphase einer solchen Infektion analysiert, welche mit einer einzelnen infizierten Zelle in einer stationären Wirtszellpopulation beginnt. Ich habe herausgefunden, dass der Infektionsprozess zu Beginn sehr erratisch verläuft, und dass die Infektion mit hoher Wahrscheinlichkeit ausstirbt. Tatsächlich geschieht dies meist schon innerhalb weniger Zellgenerationen. Meine Berechnungen haben zudem gezeigt, dass

die Infektionsprävalenz auch bei erfolgreichen Infektionen im Verlauf der Zeit nur langsam ansteigt. Weiter hat sich ergeben, dass die meisten infizierten Zellen nur wenige IS-Kopien in ihrem Genom enthalten. Dieses Ergebnis stimmt mit der von mir bestimmten IS-Verteilung in 728 vollständig sequenzierten Genomen überein. Während einer zweiten Analyse habe ich die Spätphase einer IS-Infektion untersucht. Dafür habe ich die Verteilung der IS-Kopien in den Genomen einer Wirtspopulation im Gleichgewichtszustand modelliert. Anhand dieses Modells und anhand der von mir erneut bestimmten Verteilung von IS-Kopien in 1128 vollständig sequenzierten Genomen konnte ich Maximum-Likelihood-Schätzungen für die HGT-Rate und für die Fitness-Reduktion einer IS berechnen. Es hat sich herausgestellt, dass die Fitness-Reduktion durch IS klein ist, wenn eine solche überhaupt besteht. Die geschätzte HGT-Rate entspricht etwa den Werten, welche bei in der Natur lebenden Zellpopulationen beobachtet werden. Die geschätzten Parameterwerte führten allerdings zu unrealistisch langen Zeiten bis zum Erreichen von in der Natur beobachteten Infektionsprävalenzen. Deshalb vermute ich, dass bei der raschen Ausbreitung von IS gelegentliche und vorübergehende, positive Auswirkungen auf die Fitness ihres Wirtsorganismus eine wichtige Rolle spielen könnten. In einem dritten Projekt habe ich schlussendlich die räumliche Ausbreitung von IS in einer Metapopulation simuliert. Dabei konnten die IS sowohl negative als auch positive Auswirkungen auf die Fitness ihrer Wirtszellen haben. Ich habe beobachtet, dass sowohl die Auswirkungen der IS auf die Fitness ihrer Wirtszellen als auch die HGT-Rate einen grossen Einfluss haben auf die Ausbreitungsgeschwindigkeit der IS-Infektion. Hingegen beschränkt die räumliche Anordnung der einzelnen Populationen und die Migrationsrate von Zellen zwischen den Populationen die Ausbreitungsgeschwindigkeit offenbar nicht. Meine Simulationen zeigen zudem, dass die ursprünglich infizierte Population eine entscheidende Rolle spielt in der Infektionsdynamik einer räumlich strukturierten Metapopulation.

1. Introduction

In the 1940s, Barbara McClintock discovered and described two genetic elements with peculiar properties in the maize genome, called *Ds* (short for *Dissociation*) and *Ac* (short for *Activator*) [McClintock, 1950]. *Ac* was able to autonomously excise itself from its current location in the genome and insert into another location, a process called transposition, thus silencing the expression of genes into which it might have inserted. *Ac* was also able to activate (hence its name *Activator*) the transposition of *Ds*, which could not transpose on its own. The excision of *Ds* did often lead to chromosome breakage at its current location (hence its name *Dissociation*). One of the most spectacular, phenotypically observable effects of the transposition activity of *Ac* and *Ds* was the creation of spotted maize-kernels. In some varieties of maize, one of those two genetic elements had inserted into the gene responsible for the pigmentation of kernels, which left the kernels colorless. If the element transposed in a kernel cell during cob growth, thus restoring gene expression, pigment production in that cell and all its descendants restarted, leading to a spotted kernel [McClintock, 1953]. At the time, McClintocks discovery was met with scepticism from other researchers, because it contradicted the current view of the genome as a stable collection of genes at fixed positions. Only during the 1960s, when elements similar to *Ac* and *Ds* were found in bacteria, did the idea of "mobile DNA" become popular and stimulated intense research.

1.1 Mobile DNA and transposable elements

Mobile DNA belongs to a class of genetic elements that can move within or between genomes. Viruses and plasmids represent subclasses of mobile DNA, whereas *Ac* and *Ds* represent another subclass called transposable elements. The latter can move (or transpose themselves, hence the name) from one location in the genome to another. Some elements contain all the functionality needed for their own transposition and are called autonomous (e.g. *Ac*). Their size covers a wide range of 0.7 to 20 kb [Chandler and Mahillon, 2002, Kidwell, 2002]. Some transposable elements, called nonautonomous, depend on autonomous

elements to provide the machinery for transposition (e.g. *Ds*) and can be as short as 60 bp [Tu, 2001]. An autonomous element can become nonautonomous if the gene(s) that code for its transposition functionality are inactivated through mutation or deletion. This process is so common, that most transposable elements found in a genome are in fact nonautonomous or have even deteriorated further and become inactive [Kidwell, 2002].

1.1.1 Transposition mechanism and classification

Transposable elements can be categorized into two classes, depending on their transposition mechanism [Evgen'ev, 2007,Wicker et al., 2007,Hua-Van et al., 2011]. Class I elements, also called retrotransposons, occur predominantly in eukaryotes (mobile group II introns being a notable exception) and use a copy-and-paste mechanism for transposition. Their DNA is first transcribed into an RNA intermediate, which is then reverse-transcribed into DNA by a reverse transcriptase encoded by the element itself. The newly created DNA copy is then reintegrated elsewhere in the genome, again using enzymes (e.g. integrase) produced by the element. Class II elements, also called DNA transposons, occur both in eukaryotes and in prokaryotes and often (but not always) use a cut-and-paste mechanism for transposition (*Ac* and *Ds* are examples of this class). They do not need an RNA intermediate. Instead, they often code for an enzyme, transposase, that fully extracts the element from the surrounding DNA and inserts it elsewhere in the genome. If such an element excises during cell division from a location in the genome that has already been duplicated to a location that still awaits duplication, the element may indirectly replicate itself in one daughter cell even though it only uses a cut-and-paste mechanism. Other class II elements encode for enzymes that only transfer single-stranded DNA to a new location in the genome, leaving the element DNA duplication to repair mechanisms of the host cell [Kapitonov and Jurka, 2001]. They therefore use a copy-and-paste mechanism.

1.1.2 Abundance and relevance

Transposable elements occur in the genomes of almost all organisms, although the fraction of a genome consisting of those elements may vary considerably between organisms, especially in eukaryotes. While the fraction of the genome consisting of transposable el-

ements is usually below 5% and at most 10% in prokaryotes [Hua-Van et al., 2011], the same fraction in eukaryotes is less than 5% for the yeast genome [Bleykasten-Grosshans and Neuvéglise, 2011], about 45% for the human genome [International Human Genome Sequencing Consortium, 2001], and almost 85% for the maize genome [Schnable et al., 2009]. Transposable elements reach such large fractions despite their comparatively small size because they may occur in large numbers. For example, the most frequent (and nonautonomous) element in the human genome, called Alu, has a length of only about 300 bp but occurs in roughly $1.1 \cdot 10^6$ copies in our genome, so that it nevertheless constitutes 10.6% of the human genome. [International Human Genome Sequencing Consortium, 2001]. Even members of the same genus may contain very different fractions of transposable elements in their genome. For example, that fraction varies from 39% to 66% in 8 different, diploid species of the genus *Oryza* (including rice, *Oryza sativa*), and it also roughly corresponds to genome size, in that larger genomes usually contain larger fractions of transposable elements [Zuccolo et al., 2007]. In fact, the large variation in the fraction of a genome occupied by transposable elements is one of the main explanations (besides polyploidization) for the observation that many organisms of similar complexity have strongly varying genome sizes, the so-called C-value paradox [Lynch, 2007, Chénais et al., 2012].

Transposable elements are not only abundant, they also affect their hosts. While the fraction of spontaneous mutations caused by transposable elements in humans is likely to be small [O'Donnell and Burns, 2010], a majority of the observed, spontaneous mutations in *Drosophila melanogaster* are probably caused by transposable elements [Charlesworth et al., 2004]. Obviously, the transposition process itself poses a risk for the host cell, because the element might insert into a gene, thereby deactivating it. This risk is somewhat mitigated by the fact that many transposable elements show some degree of target specificity, i.e. they have some preference as to where they integrate into the host genome [Craig, 1997]. The strictness of this target specificity is highly variable among transposable elements. While some elements insert with high probability into a unique location in the genome, where they do no harm to their host [Craig, 2002], others prefer to insert into heterochromatin or upstream of a promoter [Rio, 2002], and yet others simply use a copy of themselves as their preferred insertion target [Reimmann and Haas, 1987]. Besides directly affecting host

cell function through insertion into genes, multiple copies of a transposable element in a genome can lead to a reshuffling of the genome through ectopic recombination (recombination between copies at different locations in the genome), usually with detrimental effects to the host [Kidwell and Lisch, 2002]. The situation is made worse for hosts by the fact that transposable elements can quickly sweep through a host population under favorable conditions. For example, it took the P element, a DNA transposon, only a few decades to invade all wild populations of *Drosophila melanogaster* worldwide, starting from the Americas [Anxolabéhère et al., 1988]. But the hosts are not defenseless. Many eukaryotes have silenced most of the transposable elements in their genome through epigenetic mechanisms (e.g. methylation, histone modification, and RNA interference) [Goll and Bestor, 2005, Johnson, 2007]. In fact, some authors think that those mechanisms originated as a defense against transposable elements and similar foreign DNA integrating into the genome, and that the mechanisms were only later in their evolutionary history reused to regulate gene expression [Yoder et al., 1997, Matzke et al., 1999, Kidwell and Lisch, 2001]. In some cases, transposable elements have even been recruited by the host to fulfill a specific function. A remarkable example can be found in *Drosophila melanogaster*, where retrotransposons insert into the telomeres of each chromosome after DNA replication. Even though *Drosophila melanogaster* does not possess a telomerase reverse transcriptase, it can thus avoid telomere shortening [Kidwell and Lisch, 2001]. Transposable elements show a wide continuum of interactions with their hosts, ranging from pure parasitism to mutualism or symbiosis [Kidwell and Lisch, 2001], and they can be considered as a part of the 'ecology of the genome', a term coined by Kidwell and Lisch [Kidwell and Lisch, 1997]. The view of the genome as an ecological community has been further developed by several authors [Brookfield, 2005, Le Rouzic et al., 2007, Venner et al., 2009], who then use concepts from ecology (e.g. ecological niche and species) to formulate and analyse the interaction between different transposable elements and their host.

Because of their mobility in genomes, transposable elements represent a valuable tool for biologists and bio-engineers, especially because often one and the same element can be used over a wide range of species. For example, the presence/absence pattern of retrotransposons has been used as a phylogenetic marker [Kriegs et al., 2006, Shedlock et al., 2004]. Because

of their ability to disrupt genes, transposable elements are used for the identification of essential genes [Reznikoff and Winterberg, 2008]. In addition, transposable elements are used to deliver transgenes into host genomes, for example into mouse [Yant et al., 2000] or human cells [Yant et al., 2007], thus showing potential for future use in gene therapy.

1.2 Insertion sequences

Insertion sequences persist both in eukaryotes and in prokaryotes (bacteria and archaea) and are the simplest form of autonomous transposable elements, or more specifically, DNA transposons. I will focus on prokaryotic insertion sequences (ISs).

1.2.1 Structure and properties

ISs are short (700–2500 bp) and typically contain one open reading frame that encodes just one enzyme, transposase, which is needed for transposition [Chandler and Mahillon, 2002]. The open reading frame is usually flanked on both sides by short (10–40 bp) inverted repeat sequences that serve as recognition sites for transposase and allow for precise cleavage of the element from the surrounding DNA. In addition, the inverted repeat on the upstream end of the element contains a (usually weak) promoter for the transposase gene. The inverted repeat on the downstream end often contains (part of) an outward-directed promoter for host genes and can thus upregulate their expression. The element-specific binding domain of the transposase is often positioned in the N-terminal region of the enzyme, and the catalytic domain is often positioned in the C-terminal region. This allows the enzyme to already bind to the IS's recognition site during translation. As the fully translated catalytic domain often reduces binding activity, this arrangement of transposase domains favors enzyme activity in *cis* [Chandler and Mahillon, 2002].

Transposition can be conservative (cut-and-paste), which is the usual mode of transposition for most ISs, or replicative (copy-and-paste). During transposition, the transposase binds to the recognition sites in both inverted repeats at the ends of the IS and possibly (depending on the element) also to the target site in the host genome. The transposase then cleaves the element from the surrounding donor DNA. Repair of the double-stranded breaks of the donor DNA is left to host enzymes. The two ends of the element, still bound to the

transposase, then each nick one of the DNA strands of the target DNA, separated by an element-specific number of nucleotides (2–14 nt). The element is inserted, and the staggered ends of the target DNA are repaired by host enzymes, leading to direct repeats flanking the IS [Mizuuchi and Baker, 2002]. The length of these direct repeats is characteristic of an IS and can be used as an additional criterion during the search for ISs in published genome sequences [Wagner et al., 2007]. Some transposases combine excision from the donor DNA and insertion into the target DNA, thus creating a cointegrate, whereby both donor and target DNA are linked together by the element in a branched structure [Chandler and Mahillon, 2002]. Either the element is subsequently fully excised from the donor DNA and inserted into the target DNA, or the element is replicated by host enzymes, so that there exist afterwards two copies of the element, one at the old site in the donor DNA, and one at the new site in the target DNA. This latter process is error-prone and can lead to the deletion of host DNA, mostly with fatal consequences to the host cell [Mizuuchi and Baker, 2002]. An IS can increase its copy count even through a cut-and-paste transposition mechanism if the IS transposes during DNA replication, after the replication fork has already passed, and inserts in front of the replication fork. One of the two daughter cells will then contain two copies of the IS in its genome instead of one. For a few ISs, this is the preferred mode of transposition [Kleckner, 1989]. During the transposition process, an IS can get lost and be degraded by host enzymes (IS excision). And ISs, like all transposable elements, can also become inactive through mutations or deletions.

Many ISs show a certain degree of DNA target specificity [Chandler and Mahillon, 2002]. For example, some prefer AT-rich regions, and some are known to preferably insert into DNA sequences of the form AAA–N_{15-20}–TTT [Zerbib et al., 1985, Mayaux et al., 1984].

Two copies of an IS that flank intermediary genes and transpose synchronously, thereby mobilising the intermediary genes, constitute a composite transposon. ISs can thus be involved in transferring genes that confer resistance to antibiotics [Berg, 1989, Kleckner, 1989], genes that encode toxins [So and McCarthy, 1980], or genes with new metabolic functions [Top and Springael, 2003]. On the one hand, ISs may therefore help spreading antibiotic resistance among pathogens and can be regarded as a public health threat. On the other hand, ISs are also useful tools for genetic engineering.

Over 4000 different insertion sequences are already known and grouped into 26 families, based on (1) the primary structure of their transposase, (2) the length and sequence of their inverted repeats, (3) the length and sequence of the direct repeats generated during their insertion, (4) the organisation of their open reading frames (if they have more than one), and (5) the target sequences into which they preferably integrate (if they show some preference) [Siguier et al., 2014]. Currently, ISfinder (http://www-is.biotoul.fr) is the reference center for bacterial ISs, where reference sequences can be downloaded and new sequences can be uploaded, which will then be named according to a coherent nomenclature [Siguier et al., 2006b]. In earlier naming conventions for ISs, an IS name had the form ISn, where n was a continually increasing number, e.g. IS*1*, IS*2*, IS*3*, etc. The current naming scheme includes an abbreviation for the genus and the species in which the sequence was first found (and an additional number, if necessary), e.g. ISPpu for an IS first found in *Pseudomonas putida*.

1.2.2 Control of transposition activity (and coevolution with host)

The transposition activity of an IS, like the activity of all transposable elements, poses a potential threat to the host cell, and in consequence to the IS itself. It is therefore not surprising that ISs and their hosts have developed several mechanisms to control transposition activity [Chandler and Mahillon, 2002]. One such mechanism is the generally weak promoter of ISs and has already been mentioned above. Another control mechanism found in some ISs (e.g. IS*10* and IS*50*) protects the IS from the effects of impinging transcription, which can occur when the IS inserts into a gene and is transcribed during gene transcription. The control mechanism consists of an inverted repeat sequence located at the left end of the IS and overlapping the ribosome binding site. After impinging transcription, the inverted repeat sequences in the mRNA generate a stem-loop, thus occluding the ribosome binding site and preventing translation. Transcripts started at the IS's promoter, however, only contain one of the inverted repeats, so that no stem-loop is generated in the mRNA. A third control mechanism observed in some ISs (e.g. IS*1*, and members of the families IS*3* and IS*5*) is called programmed translational frameshifting. This process involves two open reading frames, the one upstream coding for the DNA recognition domain of the trans-

posase, and the one downstream coding for the catalytic domain. The protein product of the upstream frame by itself down-regulates transposition activity, probably by binding to the inverted repeat sequences. Only if a rare -1 frameshift occurs during translation is the catalytic domain added to the DNA recognition domain, thus generating a functional transposase molecule. A fourth control mechanism has to do with the sensitivity of transposases to host enzymes and temperature, so that transposases usually have a short half-life. As already mentioned, many transposases can bind to their IS's recognition site even during translation, because the binding domain of the enzyme is commonly positioned in the C-terminal region. Both short half-life of the transposase and the possibility of DNA binding during translation lead to preferential activity of the transposase in *cis* and reduce the risk of transposition through the transposase expressed by another copy of the IS elsewhere in the genome [Chandler and Mahillon, 2002].

ISs and their hosts have co-evolved over a long time. Hosts have therefore developed their own control mechanisms of transposition activity. Among other mechanisms, DNA chaperones play a role in transposition control. For example, several ISs contain specific IHF binding sites within or near the transposase promoter, and it has been shown that IHF can down- or upregulate the activity of Tn*10* [Chalmers et al., 1998], a composite transposon that is flanked by IS*10*. Furthermore, the Dam DNA methylase can often control the expression and activity of transposase, because many ISs carry methylation sites in their promoter regions or in their inverted repeats [Chandler and Mahillon, 2002].

Taken together, the mechanisms of transposition suppression by the hosts and by the ISs themselves are usually very effective and lead to low transposition rates between 10^{-7} and 10^{-4} events per IS element and cell generation [Kleckner, 1989, Chandler and Mahillon, 2002, Sousa et al., 2013].

1.2.3 Horizontal gene transfer

Prokaryotes are asexual and have large effective population sizes. Predominantly detrimental ISs could therefore not invade a host cell population, were it not for horizontal gene transfer (HGT) [Hickey, 1992, Lynch and Conery, 2003]. HGT encompasses three mechanisms through which foreign DNA can be transferred between different cells and be

incorporated into a prokaryote's genome: (1) through direct uptake of extracellular DNA from the surrounding medium (transformation), (2) through plasmids (conjugation), and (3) through bacteriophages (transduction) [Davison, 1999]. I will now briefly discuss those three HGT mechanisms.

Transformation: In natural environments, extracellular DNA is abundant, either released from lysed cells or secreted by living cells. Extracellular DNA in contact with water usually gets quickly degraded and has a half-life of only a few hours, at least if it is not adsorbed on particulate matter in freshwater or in soil [Lorenz and Wackernagel, 1994]. Many prokaryote species are naturally competent, i.e. they can take up extracellular DNA of a few dozen kb from the environment and integrate it in their genome under specific conditions (e.g. depending on growth conditions or cell density) [Lorenz and Wackernagel, 1994, Coupat et al., 2008]. During DNA uptake, double- or single-stranded DNA is bound to the cell surface, processed through the outer membrane (for gram-negative bacteria) and the plasma membrane, and released as single-stranded DNA into the cytoplasm. There it is either degraded or integrated into the genome, depending on the existence and the length of a homologous sequence on the single-stranded DNA and the host genome [Lorenz and Wackernagel, 1994, Chen and Dubnau, 2004, Thomas and Nielsen, 2005].

Conjugation: Plasmids are extra-chromosomal and usually circular DNA sequences that can replicate independently of their host cell. Some plasmids, called conjugative plasmids, code for additional functionality that allows them to copy themselves and get transferred through a pilus (tube-like membrane) to a neighbouring cell [Norman et al., 2009]. Conjugative plasmids have a size between 30 kb and several hundred kb. They often carry additional genes (for example encoding antibiotic resistance, pathogenicity, or new metabolic pathways) and transposable elements, among them ISs, which can transpose and insert into the host genome. Because of their large size, conjugative plasmids are usually present only in low numbers in a cell. To make sure that only daughter cells survive that contain at least one plasmid copy after cell division, conjugative plasmids often implement a post-segregational killing system consisting of a long-lived toxin and a short-lived antidote. If a daughter cell does not contain the plasmid after cell division, the short-lived antidote decays, and the toxin kills the cell [Bahl et al., 2009].

Transduction: Bacteriophages (phages) encapsulate their DNA or RNA in a protective protein shell (capsid). They have a half-life of several days in seawater and soil, but under good conditions, phages can stay infective for years [Bratbak et al., 1994, Weinbauer, 2004]. Phages can show basically two different life cycles, a lytic and a lysogenic one. In the lytic cycle, after entering the cytoplasm, a (virulent) phage induces the cell to produce new phages, which are released to the environment after cell lysis. In the lysogenic cycle, a (temperate) phage inserts into the genome of the host cell after entering the cytoplasm and enters a dormant state (prophage). Under certain conditions (e.g. if the cell is stressed), a prophage excises from its surrounding DNA into the cytoplasm and enters the lytic cycle. Phages have a genome size of 3–300 kb [Hatfull, 2008]. They are abundant and usually outnumber prokaryotic cells in almost any environment where prokaryotes live. Furthermore, a large proportion of prokaryotes is infected with phages. It has been estimated that about 35% of all prokaryotes in marine environments are infected with a functional viral genome [Weinbauer, 2004].

While there are thus different ways in which foreign DNA can enter a cell via HGT and insert into its genome, there also exist several barriers to HGT. It has already been mentioned that the integration of extracellular DNA into a genome depends on homologous sequences on the DNA, i.e. integration occurs preferentially into the DNA of closely related prokaryotes, and unrelated DNA will usually be degraded by restriction endonucleases [Thomas and Nielsen, 2005]. The same can happen to a freshly acquired plasmid. Some plasmids have adapted to restriction by reducing the number of restriction sites, or by encoding proteins that interfere with the host's restriction mechanism [Thomas and Nielsen, 2005]. The CRISPR-Cas system (CRISPR is a shortcut for 'clustered regularly interspaced short palindromic repeats') is a prokaryotic defense mechanism that is mainly directed against phages and plasmids. About 45% of all known bacteria and 85% of all known archaea use such a system [Bondy-Denomy and Davidson, 2014]. It consists of an array of alternating repeats (21–48 bp) with interspersed spacers (26–72 bp). The spacers are pieces of foreign DNA that has been encountered earlier. The transcript of such a spacer serves as a recognition site and together with a Cas protein builds a complex that surveys the cell for foreign DNA.

ISs can in principle infect a host genome through all three forms of HGT, i.e. transformation, conjugation or transduction. The integration of an IS through transformation and transduction could happen by chance, either because a piece of extracellular DNA contains an IS, or because a phage contains a piece of host DNA with an IS that had been accidentally integrated during phage production. However, more probable is an infection through plasmids. It has been observed that large plasmids, such as conjugative plasmids, contain a comparatively large number of ISs, while only few phages contain ISs [Siguier et al., 2006a, Leclercq and Cordaux, 2011]. This has been explained by purifying selection acting stronger on phages than on plasmids, because phages have a lower proportion of intergenic sequence and at the same time a higher proportion of essential genes than plasmids [Leclercq and Cordaux, 2011]. The importance of plasmids in exchanging genetic material (including ISs) even between different species has also been shown in a network analysis encompassing cellular genomes and plasmid and phage DNA sequences [Halary et al., 2010]. In general, rates of HGT events through transformation, conjugation and transduction depend strongly on circumstances (e.g. free-living bacteria in water or soil, or bacteria living in biofilm) and cover a wide range of values. Table 1.1 shows several rates given in the literature.

HGT event	Rates	Sources
Transformation	10^{-6} – 10^{-3}	[Williams et al., 1996]
Conjugation	10^{-6} – 10^{-5}	[Dahlberg et al., 1998]
Transduction	10^{-8}	[Jiang and Paul, 1998]

Table 1.1: HGT rates reported by different authors. Rates have been converted to numbers of events per cell and generation.

HGT is common among prokaryotes, and because it shuttles genes between prokaryotic species that live in close proximity but are not necessarily related, it blurs the phylogenetic relationship between those species [Gogarten and Townsend, 2005, Kloesges et al., 2011]. For example, in three different, fully sequenced strains of *Escherichia coli*, only about 40% of the combined, nonredundant set of proteins are common to all strains [Welch et al., 2002]. HGT may also pose a public health problem, because it helps spreading antibiotic resistance and virulence genes [Aminov, 2011, Juhas, 2015].

1.2.4 Distribution of insertion sequence counts

Transposable elements are much less frequent in prokaryotes than in eukaryotes, which is usually attributed to the prokaryote hosts' smaller genome sizes and the stronger selection pressure on this size in prokaryotes [Hua-Van et al., 2011]. Even within prokaryotes, there is a correlation between genome size and IS count [Touchon and Rocha, 2007]. In fact, larger genomes do not only have higher IS counts, but the IS density is higher, too. This has been explained based on the following two observations: (1) about 90% of a prokaryotic genome consists of genes, and much of the rest has a regulatory function, (2) prokaryotes have a set of about 500 genes that are essential and/or ubiquitous. In a small genome, a large DNA fraction therefore consists of essential or ubiquitous genes, and an IS is much more likely to disrupt one of these genes than in a large genome, which may contain more genes of less importance to the survival of the cell [Touchon and Rocha, 2007]. In figure 1.1, I show the distribution of the number of ISs per genome (IS count) of the six most abundant ISs, which have infected at least 30 of the 1128 fully sequenced genomes I considered. Figure

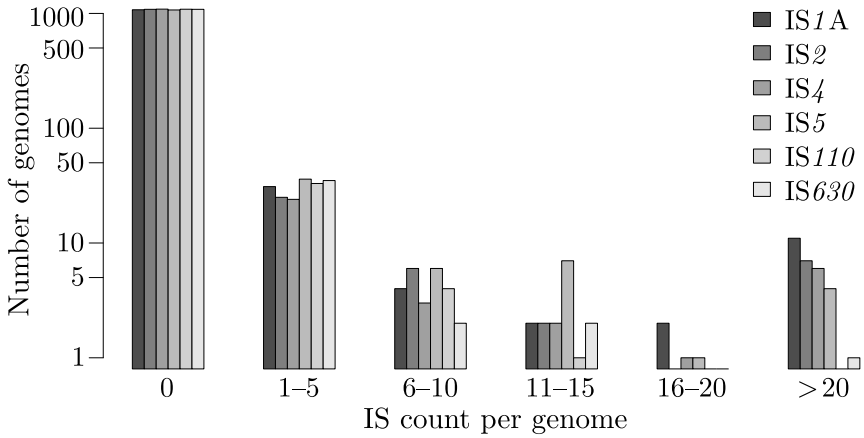

Figure 1.1: IS count distribution of six ISs that have infected at least 30 of 1128 fully sequenced genomes (as of July 2010). The genome sequences are from the National Center for Biotechnology Information, NCBI [NCBI, 2010], and the reference sequences of the ISs are from the ISfinder database [Siguier et al., 2006b]. ISs in the genomes have been identified and counted using IScan [Wagner et al., 2007]. Note the logarithmic scale on the vertical axis.

1.1 shows that the IS count distribution for each of the six ISs is strongly L-shaped: a large majority of the genomes contain no IS copy, a small fraction of genomes contains up to 5 or 10 copies, and only few genomes contain more than 10 copies, although there are

genomes with much higher IS counts (e.g. a maximum of 220 copies of IS*1*A in *Shigella boydii* CDC 3083-94). In a similar analysis, two authors reported in 2007 that 24% of 262 fully sequenced genomes contained none of the IS elements known at that time [Touchon and Rocha, 2007]. This indicates that ISs in general have a detrimental effect on their host's fitness. Together with the observation that ISs inside the same host genome show high nucleotide similarity, while ISs between different genomes do not [Wagner, 2006], this makes the following evolutionary scenario plausible: an IS spreads through a host cell population, perhaps aided by possible beneficial side effects (e.g. because it is part of a composite transposon), and then quickly increases its copy number per genome by transposition. However, with increasing copy number per genome, selection against the IS gets stronger, and over time, the element (and perhaps also the bacterial strain carrying it) becomes extinct. The infection-extinction cycle may then repeat itself [Wagner, 2006, Siguier et al., 2006a]. A similar cycle has been proposed for DNA transposons in eukaryotes [Brookfield, 2005], based on a model of the population dynamics of the P element [Brookfield, 1991]. In that model, a transposase entering the nucleus acts in *trans*, which favors the spreading of nonautonomous elements that can repress transposition. If transposition is repressed, there will in turn be no selection to maintain the sequence encoding transposase, i.e. autonomous elements will die out. After even the repressing nonautonomous elements have decayed, the species will again be susceptible to an invasion of a similar element.

1.3 Population dynamics of transposable elements

Transposable elements are selfish DNA sequences, in that they encode only what they need for their own replication [Doolittle and Sapienza, 1980, Orgel and Crick, 1980]. The relationship between transposable element and host is therefore basically a relationship between parasite and host. Therefore the population dynamics of transposable elements can be analysed with the help of epidemic models. This does not mean that transposable elements cannot have positive effects on their hosts (and in fact they do have such effects occasionally or temporarily), but those effects are not necessary for the persistence of transposable elements and may be consequences rather than causes of their presence [Ajioka and Hartl, 1989, Charlesworth et al., 1994]. There is another reason to assume that transpos-

able elements are genomic parasites. If they had mostly beneficial effects on their hosts, the elements should have gone to fixation at the locations where such effects occur. But this has not been observed [Montgomery and Langley, 1983, Sawyer et al., 1987, Ewing and Kazazian, 2010]. Therefore the occasionally beneficial effects of transposable elements cannot be the main cause of their presence.

1.3.1 Earlier population models for transposable elements

Since their discovery by McClintock, transposable elements had been puzzling researchers, because with few exceptions, it had proven difficult to assign an adaptive role in the genome to those elements. Two articles by Doolittle and Sapienza [Doolittle and Sapienza, 1980] and by Orgel and Crick [Orgel and Crick, 1980] marked a turning point in how researchers thought about transposable elements. Those articles introduced the concept of selfish DNA, i.e. genetic elements that do not necessarily fulfill an adaptive function for their host but exist as genomic parasites. After publication of these articles, many studies extending this perspective appeared. I will now present a few of those studies.

If transposable elements are selfish DNA sequences, with generally detrimental effects on their hosts, then why do they persist? Hickey gave an answer to this question for a sexually outbreeding population of eukaryotic hosts [Hickey, 1982]. He showed that even an element with strongly detrimental fitness effects could spread through a population, provided that the replicative transposition rate of the element was high enough. His argument went as follows: assume that a neutral element's replicative transposition rate were high enough, so that almost all gametes of a parent whose zygote contained only one element would contain at least one element. Then the element would initially almost double its frequency in the population during each generation. Or, interpreted in another way, it could spread in the population even if it reduced the fitness of its host by almost 50%, i.e. if it had a strongly detrimental effect on its host's fitness. The argument is of course simplistic, and the situation described above may be extreme, but nevertheless, the argument shows that transposable elements do not have to fulfill an adaptive function for the host to persist. On the contrary, Hickey's reasoning leads to the somewhat counter-intuitive insight that detrimental transposable elements can invade a host population and actually *decrease*

the population's mean fitness. This is not purely hypothetical, as it has been shown in the laboratory that the frequency of genomes carrying the P element, a DNA transposon in *Drosophila melanogaster*, can in fact double every generation, thus explaining the fast spreading of this element through all wild populations worldwide during only a few decades, despite the element's detrimental effects on its host. Because transposable elements need a form of sex to spread in a population, Hickey speculated that the elements themselves might have been a driving force in the invention of sex [Hickey, 1982, Hickey, 1992]. This line of thought is continued by Johnson and Brookfield, who examine the fate of a selfishly spreading gene that increases the probability of its diploid host to reproduce sexually (many single-celled eukaryotes are facultative sexuals, and sex occurs only occasionally) [Johnson and Brookfield, 2002]. In their model, both the gene and sexual reproduction carry a cost. The authors assume that if it occurs, mating is random, and parents that are initially homozygous or heterozygous for the gene have offspring that always carries the gene. The authors show that these assumptions can lead to complex population dynamics, depending on parameters: the gene may get lost, may go to fixation, a stable or an unstable, non-trivial equilibrium of carriers and element-free hosts may exist, or the gene may be effectively neutral. The authors show that in general, if a selfishly spreading gene increases its carrier's frequency of sex, it increases the probability to successfully invade a host population, but at the same time decreases the probability to go to fixation once it has established itself in the host population.

Assuming that the basic transposition rate per element is higher than its excision rate, an equilibrium in the number of transposable elements per genome can only be reached if the transposition rate per element decreases for increasing copy number (autoregulation), if the excision rate increases with increasing copy number, or if higher copy numbers cause negative fitness effects so that selection acts against an unlimited increase in the number of transposable elements per genome. Excision rates are at least an order of magnitude smaller than transposition rates [Kleckner, 1989, Sousa et al., 2013, Maside et al., 2000], and an increase in their rate with increasing element copy number has not been observed. Therefore, research on the population dynamics of transposable elements focuses on autoregulation and on selection against increasing copy numbers [Ajioka and Hartl, 1989].

Charlesworth and Charlesworth examined the effects of autoregulation in transposition and of selection against increasing copy numbers of transposable elements in randomly mating populations [Charlesworth and Charlesworth, 1983]. The authors were interested in conditions under which an equilibrium in the element count can exist, using different analytical and simulation models. They found that indeed both autoregulation and selection can lead to an equilibrium in the element count per genome. In the case of selection against high element copy numbers without autoregulation, the authors found that not just any decrease of host fitness with increasing copy number will necessarily lead to an equilibrium in the copy number. For a population with infinite size, the graph of host fitness w_n versus copy number n must be downward curved (decreasing and strictly concave), i.e. w_n must decrease stronger than linearly with increasing copy number n. This also means that the detrimental fitness effects s_d of each of the n elements in a genome cannot be independent, which would lead to a fitness function of the form $w_n = (1 - s_d)^n$, and neither can the detrimental fitness effects be additive, which would lead to a fitness function of the form $w_n = 1 - ns_d$. For a suitable fitness function w_n, the mean count $\bar{n}$ of transposable elements per genome is then defined by $|\partial \ln (w_{\bar{n}}) / \partial \bar{n}| \approx u - v$, where u is the transposition rate and v is the excision rate (both rates are constant and given per element and generation).

Langley et al. studied a model for the spreading of a transposable element without fitness effects in a finite Mendelian population [Langley et al., 1983]. They assumed autoregulation in transposition, i.e. a decreasing transposition rate per element with increasing element count, but no regulation in the loss of elements. The authors derived an expression for the frequency spectrum of the infection process at equilibrium, which allows one to calculate for example the expected number of locations in a haploid genome at which at least one individual in the population contains a transposable element. The expression depends on two parameters, one of which is the expected (or mean) number of transposable elements per haploid genome at equilibrium. Kaplan and Brookfield developed a method to estimate these two parameters based on data. They then applied their method on data from Montgomery and Langley about the distribution of *copia*-like retrotransposons in *Drosophila melanogaster* [Montgomery and Langley, 1983], to estimate the two parameters in the expression for the frequency spectrum provided by Langley et al. [Kaplan and Brookfield,

1983], which in turn lead to good model fits for data from *Drosophila melanogaster*.

Kaplan et al. extended the model of Langley et al. and included the possibility that an autonomous, wild-type transposable element can mutate to a nonautonomous element that can only transpose if the wild-type element is present in the genome [Kaplan et al., 1985]. The authors show that transposable elements then go extinct. There are two ways to extinction for a newly introduced, autonomous element: either the element cannot establish itself and vanishes shortly after introduction, or the element can establish itself in the population, reaches a quasi-equilibrium that may persist for a long time, while nonautonomous, mutated elements slowly replace autonomous, wild-type elements and ultimately get lost themselves. The authors predict that in *Drosophila melanogaster*, specific retrotransposons, which show low variability in their sequence, and therefore seem to have a small mutation rate and thus will generate nonautonomous elements at a small rate, will stay in the genomes of their hosts for a long time. DNA transposons, on the other hand, which usually show more sequence divergence, and seem to have a high mutation rate, thus generate nonautonomous elements at a high rate and will have a short, less stable evolutionary history.

Even a purely detrimental transposable element can spread in a sexually propagating host population [Hickey, 1982]. The spread of an element in (almost) clonal prokaryotes is more difficult to explain (apart from the fact that prokaryotes have higher effective population sizes and thus stronger selection against useless, selfish DNA [Lynch and Conery, 2003]). The only means by which an element can infect another prokaryote host cell is HGT. Because HGT rates are generally low (see table 1.1), Condit et al. argued against the hypothesis that transposable elements in prokaryotes are generally parasitic DNA sequences [Condit et al., 1988]. Instead, they favored the hypothesis that those elements are maintained in populations by increasing the fitness of their hosts. In their model, they used plasmids as vectors for transposable elements in a host cell culture assumed to be at constant concentration in a chemostat. Only one element per genome or plasmid was allowed, and the element could transpose (conservatively or replicatively) from the plasmid to the genome and vice versa. The authors then used a system of ordinary differential equations to calculate the densities of cells with and without plasmids, as well as with and without transposable element inserted in the plasmid or in the genome.

To study the fate of a single transposable element introduced into an uninfected host population, branching process models (see also subsection 1.4.1) are useful and popular. Multi-type branching processes, where the type indicates the number of transposable elements per genome, allow in a natural way to model the distribution of the number of those elements per genome in an infected population [Ajioka and Hartl, 1989].

Sawyer et al. studied the count distribution of six ISs (IS*1*, IS*2*, IS*3*, IS*4*, IS*5*, and IS*30*) from five different families (IS*1* and IS*2* belong to the same family) in a reference collection of 71 strains of *Escherichia coli* from humans and animals and from diverse geographic locations (ECOR collection) [Sawyer et al., 1987]. Using a continuous-time multi-type branching process model, the authors used the data to select from a given set of functions the ones that fitted transposition regulation and fitness in dependence of IS number per genome best. In their model, the type of the multi-type branching process corresponded to the copy number per genome of a specific IS. The authors neglected excision in their model and assumed that the transposition rate $T = T(n)$ *per genome* in dependence of the number n of ISs per genome had the form $T(n) = T_0\, n^i$ with $i \in \{-1, -1/2, 0, 1/2, 1, 2\}$, so that for $i < 1$, the transposition rate *per IS copy* would be down-regulated for increasing IS count. They assumed the cell division rate R to be constant, but modeled the death rate $D = D(n)$ of a host cell containing n ISs in its genome in a similar way as the transposition rate, namely $D(n) = D_0\, n^j$ with $j \in \{-1, -1/2, 0, 1/2, 1, 2\}$. The authors defined the fitness function as $R - D(n)$. They found that none of the six distributions of IS numbers per genome allowed an unequivocal choice among the best-fitting functions for transposition regulation and for fitness. Instead, different combinations of transposition regulation and fitness function showed similarly good fits for different ISs. The authors found that the only model acceptable for all six ISs showed a moderate regulation of transposition ($i = 0$) and a weak effect of copy number on fitness reduction ($j = 1/2$).

Basten and Moody added selection depending on IS numbers per genome to a discrete-time multi-type branching process model developed earlier by Moody [Moody, 1988], and used this extended model to calculate equilibrium distributions of IS numbers per genome, depending on different selection functions [Basten and Moody, 1991]. In their model, only one transposition or excision event can happen during one generation, and both the (replica-

tive) transposition and the excision rates per genome and generation are linear functions of the IS number per genome. The former is decreasing (autoregulation) and the latter increasing. Selection against high copy numbers is modeled as follows: the expected number of offspring per cell after one generation is larger than zero for uninfected cells, has a constant value (smaller or larger than for uninfected cells, depending on whether the IS is detrimental or beneficial) from one IS copy up to a threshold number of l copies per genome, and it decreases linearly with a further increase in the IS number. In addition, the authors assumed that HGT generates newly infected cells with one IS copy in their genome at a constant rate, independent of the number of already infected cells. The authors found that an IS with detrimental fitness effects larger than the HGT rate cannot invade the host population and exists only at low frequencies dictated by the HGT rate. ISs whose fitness cost is not larger than the HGT rate can invade the host population. The frequency distribution of IS counts in equilibrium is unimodal, with the mode at $\min(l, n_0)$, where n_0 denotes the IS number at which the transposition and the excision rate per genome are equal.

A characteristic of many modeling approaches for transposable elements is that the environment is assumed to be constant, and an equilibrium distribution of element counts is calculated for given parameter sets. In the wild, however, environments are changing, and many parameters affecting the spread of transposable elements may vary over time. For example, an element that has inserted at a specific location in the host genome may be beneficial in one environment and detrimental in another, e.g. because it has inserted into a gene whose expression may cause a metabolic burden in one environment and may be necessary in another. Edwards and Brookfield devised a model reflecting such a situation [Edwards and Brookfield, 2003]. In their model, they assumed a bacterial population infected with ISs, which all cause the same basic fitness cost per element. The population lives in two different, alternating environments. In the first environment, IS insertion at a specific location of interest (e.g. a gene) is favored, and in the second environment it is disfavored, both in addition to the basic fitness cost. The authors also assumed the existence of neutral locations, where an insertion never causes any additional benefit or cost, and detrimental locations, where an insertion always causes an additional fitness cost. Transposition

in the model is replicative, occurs with a constant rate per IS copy, and leads with specific probabilities to the insertion of an IS at one of the three types of genome location described above. The authors modeled horizontal transmission of ISs so that a certain proportion of transpositions generate insertions in randomly chosen cells. IS excision is precise (i.e. leaves no remnant of the IS behind) and occurs with constant rate per IS. The authors found that an IS meeting these assumptions cannot be maintained in a population without favored insertions in the first environment. But for a wide range of parameters, including the time length of the environment's cycle, the IS could persist in the population, albeit at varying frequencies during an environment cycle. The first environment would favor those cells with high IS copy numbers at neutral sites, because those cells have a higher probability of an insertion at the location of interest. While in the second environment, insertions at the location of interest would get lost, but enough insertions at neutral locations would persist, so that the cycle could be repeated. Interestingly, HGT decreases the probability that the IS is maintained in the host population, because it disturbes the linkage between the location of interest and neutral locations.

1.3.2 Spatial population models

The infection of a host population by transposable elements is not only a process happening in time (i.e. the prevalence of infected organisms in the population varies over time) but also in space (i.e. infected members of the population show a spatial pattern of element abundance at any point in time). It is therefore useful to consider spatial models of population dynamics in examining the spread of a transposable element. Such spatial models are already used in ecology and epidemiology, and I will now present some of them, as background to the simulation model used in chapter 4.

In spatial models, a population is often split into distinct subpopulations, usually with strong interaction between individuals within each subpopulation and weaker interaction between different subpopulations (e.g. migration). Such a collection of subpopulations, together with their interactions, is sometimes called a metapopulation. A basic distinction can be made between spatially implicit and spatially explicit models. While the subpopulations in a spatially implicit model are not located in space, each subpopulation in a

spatially explicit model has a specific location in space. Furthermore, all subpopulations in spatially implicit models are often equally well connected to each other (e.g. migration rates between all subpopulations are the same), but in spatially explicit models, the connectedness between two subpopulations usually depends on the distance between them [Hanski, 1999, p. 13]. Spatially explicit models are attractive because they allow realistic modeling, but unfortunately, they are also more difficult to analyse than spatially implicit models.

One of the most simple yet still useful (spatially implicit) models from ecology is the Levins metapopulation model [Levins, 1969]. Assume that there is a large number of habitat patches that can be populated or empty (for epidemiological purposes, replace 'populated' by 'infected' and 'empty' by 'uninfected'). Migrants from populated patches can populate any empty patch with the same probability. Let $P = P(t)$ be the fraction of patches that is populated at time t. The migrants colonize empty patches in proportion to the fraction $(1 - P)$ of empty patches, and populated patches can become empty again (i.e. their population can go extinct). These assumptions lead to the differential equation

$$\dot{P} = \frac{dP}{dt} = cP(1 - P) - eP = (c - e)P\left(1 - \frac{P}{1 - e/c}\right),$$

where c and e are the colonization and extinction parameters, respectively [Hanski, 1999, p. 55]. The expression after the last equal sign tells us that the colonization (or infection) process of a set of empty patches will proceed in an analogous way to the logistic growth of a single population with an initial growth rate $c - e$ and 'carrying capacity' $1 - e/c$, which in this metapopulation model equals the stable equilibrium fraction of populated (or infected) patches (in addition to the trivial equilibrium at $P = 0$). As long as $e > 0$, a certain fraction of patches will always be (temporarily) empty, however strong the colonization force represented by c is. This observation highlights an important result from metapopulation theory: the extinction of local (sub)populations is a normal and recurrent process, and even a metapopulation consisting of locally unstable subpopulations that are prone to extinction (high e) can persist if it also has a strong colonization force (high c) [Hanski, 1998].

When the Levins metapopulation model is used in an epidemiological context, patches are either uninfected or fully infected. This assumption is justified if the dynamics between patches is much slower than the dynamics within patches, i.e. when migration rates between

patches are comparatively low. If the migration rates become larger, infected individuals that immigrate can have a considerable effect on the infection dynamics within a patch. The dynamics within a patch then needs to be modeled in more detail. A possible approach is to extend a single-population SIS model to several populations. In an SIS model (*S*usceptible-*I*nfected-*S*usceptible), S and I denote the proportions of susceptible and infected individuals, respectively, in a large, well-mixed population. A susceptible individual can get infected, will then recover from the infection and become immediately susceptible again. To describe the change in S and I over time through a system of ordinary differential equations, we need the *force of infection*, i.e. the rate of new infections *per capita* of susceptible individuals. If infected and susceptible individuals mix freely, the force of infection is βI (density-dependence or mass action [McCallum et al., 2001]), where β is the so-called transmission coefficient. Neglecting birth and death (and differences in the birth and death rates between susceptibles and infected), we can therefore describe the infection dynamics with the equation system

$$\begin{aligned} \dot{S} &= \gamma I - \beta SI \\ \dot{I} &= \beta SI - \gamma I, \end{aligned} \tag{1.1}$$

where γ is the recovery rate of the infection. This simple model already has some interesting properties. For example, there is an infection threshold, i.e. the infection can only succeed if $S > \gamma/\beta$ holds at the beginning of an infection. The inverse of this threshold is the basic reproductive ratio $R_0 = \beta/\gamma$, which corresponds to the mean number of secondary infections caused by a single infected individual during its infective period in an entirely susceptible population. It is intuitively clear that for an infection to succeed, $R_0 > 1$ is necessary. Because $S + I = 1$, usually only the differential equation for the infected population is written down, and the system (1.1) resumes to

$$\dot{I} = \beta SI - \gamma I = (\beta - \beta I - \gamma)I = \beta I\left((1 - 1/R_0) - I\right).$$

A nontrivial equilibrium for the proportion of infected individuals then exists at $I^* = 1 - 1/R_0$, i.e. for $R_0 > 0$ the infection becomes endemic [Keeling and Rohani, 2008a, p. 19f

and 39].

The single-population SIS model can be extended to an SIS model for a metapopulation with $n > 1$ subpopulations. For simplicity, we assume that all subpopulations have the same, constant size over time, and that migration from subpopulation j to subpopulation i occurs with a constant rate m_{ij}. Thus, again neglecting birth and death, this leads to the following system of ordinary differential equations for the proportion I_i of infected individuals in subpopulation i:

$$\dot{I}_i = \beta S_i I_i - \gamma I_i + \sum_{j=1}^{n} m_{ij} I_j - \sum_{j=1}^{n} m_{ji} I_i \quad (1 \leq i \leq n).$$

The migration rate matrix $M = (m_{ij})_{0 \leq i,j \leq n}$ describes the coupling between different subpopulations. The stronger the coupling is (i.e. the higher migration rates are), the more synchronous is the infection dynamics in different subpopulations, and the metapopulation then behaves almost like a large, well-mixed population. The migration rate matrix M may reflect the spatial structure of the metapopulation, in that m_{ij} may depend on the distance between two subpopulations i and j. In addition, M need not be symmetric, so that one can model source-sink-dynamics, where an infection in a subpopulation (sink) can only persist because of the constant immigration of infected individuals from surrounding subpopulations (source) [Hanski, 1999, p. 50ff]. The infection dynamics of the metapopulation may be computed numerically or, with some simplifying assumptions (e.g. about the coupling), analytically.

To analyse the spreading of an infection on a plane (or in three-dimensional space), one can use the approximation of partitioning the plane (or the space) into a grid of discrete and neighboring subpopulations. This leads to coupled lattice or grid models [Keeling and Rohani, 2008a, p. 255f]. Popular choices for partitioning a plane are square lattices or, if one wants to attenuate their strong orientation in four directions, hexagonal lattices. In a coupled lattice model, coupling is usually restricted to nearest neighbors, but can be extended to subpopulations further away. If coupling is constant and with nearest neighbors only, i.e. if a constant migration rate m to nearest neighbours is assumed, an SIS model

with equal subpopulation sizes and without birth and death has the form

$$\dot{I}_i = \beta S_i I_i - \gamma I_i + \sum_{j=1}^{n} m_{ij} I_j - \sum_{j=1}^{n} m_{ji} I_i \quad (1 \leq i \leq n)$$

$$\text{with } m_{ij} = m_{ji} = \begin{cases} m & \text{if subpopulations } i \text{ and } j \text{ are neighbors} \\ 0 & \text{otherwise.} \end{cases}$$

To avoid spurious dynamical effects at the lattice boundaries (resulting from a boundary subpopulation not being fully surrounded by other subpopulations), opposing boundaries are in practice often connected with each other, which for an initially rectangular grid results in a torus.

All of the population and infection models described above (spatial or not) have their stochastic counterparts, where probabilities are used instead of differential equations. Usually, stochastic infection models show a higher infection threshold and lead to slower infection spread than in the corresponding deterministic models [Keeling and Rohani, 2008a, p. 238f].

The coupling between subpopulations in a metapopulation or in a coupled lattice is crucial for the spatial population or infection dynamics. In a spatial model for the spreading of an invasive agent (think of an invasive plant species or of an infection), the coupling depends on how many offspring a parent produces and how far from the parent they disperse. The latter is described by the so-called dispersal kernel, the probability density function of the distance between parent and offspring. The speed with which an invading population spreads through a spatial habitat is mainly determined by the tail of the dispersal kernel, i.e. by the probability of long-distance dispersal [Kot et al., 1996]. As an illustration, assume that the invasion of a large, uninhabited territory starts from a point source. If the tail of the dispersal kernel is bound by an exponential function, the invasion of the uninhabited territory will proceed approximately in the form of an ever expanding disk of populated territory. There will be a well-defined invasion wave front between unpopulated and populated territory, and this wave front will expand with constant velocity. If, on the other hand, the tail of the dispersal kernel cannot be bound by an exponential function (e.g. if the tail follows a power law function), the invasion of the uninhabited territory

will proceed much more irregularly. There will be individuals that jump far ahead of the bulk of the population and may start their own invasion, which will later merge with the main invasion, i.e. there will be no well-defined wave front. Moreover, the speed of the invasion, defined by the square root of the populated area, divided by time, increases over time [Mollison, 1972, Kot et al., 1996, Lewis and Pacala, 2000], i.e. the invasion proceeds faster and faster.

1.4 Computational and mathematical tools

In my analysis of the population dynamics of ISs, I used various computational and mathematical tools. The main tool in chapter 2 is the branching process, which I use to investigate the early phase of an IS infection. In chapter 4, I use the tau-leaping algorithm to simulate the IS infection dynamics inside a subpopulation of host cells. Because these tools were central to my work, I describe them in the following two subsections.

1.4.1 Branching Processes

The population and infection models described in subsection 1.3.2 have been deterministic, in that the outcome of a particular population or infection process is strictly determined by its parameters and will always be the same. Such invariance is not observed in the wild or even in the laboratory, where experimental conditions can be controlled. There, chance events play an important role, and population or infection processes may have different outcomes even when they start under identical conditions (one infection may succeed, and another one might not). Branching processes are a class of stochastic processes and are well-suited to deal with the randomness that can for example occur at the beginning of a population growth process or of an infection process, when there exist only few (infected) individuals who are strongly affected by chance events (see chapter 2).

The simplest type of branching process is the single-type (Bienaymé-)Galton-Watson process, which is a single-type, discrete-time branching process [Haccou et al., 2005, p. 13]. In this process, all individuals are of the same type and have the same life span, for simplicity assumed to be one generation, after which they have k offspring ($k \in \mathbb{N}$) with probability p_k, independently of each other. Each individual having the same life-span

distribution and the same probability distribution $\{p_k\}_{k\in\mathbb{N}}$ for k offspring, independently of each other, is in fact the defining property of branching processes. Questions one may then ask include "What is the probability of the population to die out?" or "Will the population grow, and how fast will it grow?". Observe that in a deterministic model (see subsection 1.3.2), the first question is meaningless: there is no randomness involved and therefore no *probability* of dying out.

Next, we'll answer these two questions for a simple example of a single-type Galton-Watson process. Assume that a cell lives for a specific amount of time (one generation) and then either dies ($k = 0$) with probability $p_0 = 0.2$ or divides ($k = 2$) with probability $p_2 = 0.8$. The progeny of the cell (and the progeny of the progeny, etc.) then again lives for the same amount of time and dies or divides with the same probabilities. What happens to the population process starting with one cell? It is intuitively clear that the mean number m of offspring per cell will be important. If $m < 1$, $m = 1$ or $m > 1$, the single-type Galton-Watson process is called subcritical, critical or supercritical, respectively. On average, supercritical processes show geometric growth, subcritical processes show geometric decay, and critical processes show no growth nor decay. In our example, $m = \sum_{k\in\mathbb{N}} p_k k = p_0 \cdot 0 + p_2 \cdot 2 = 1.6 > 1$, i.e. the process is supercritical, and the population will on average grow geometrically. Nevertheless, the population has a non-vanishing risk of dying out by chance, especially as long as the population count is still low. To calculate the extinction probability of a single-type Galton-Watson process, one can use its generating function. The generating function f of a single-type Galton-Watson process with probabilities p_k for k offspring is defined as $f(s) = \sum_{k\in\mathbb{N}} p_k s^k$, i.e. it is a power series with coefficients p_k [Athreya and Ney, 1972, p. 2]. In our example, the generating function is $f(s) = 0.2 + 0.8s^2$. It has been shown that the extinction probability q of the single-type Galton-Watson process is the smallest non-negative solution of the equation $f(s) = s$ (observe that $s = 1$ is always a solution because $f(1) = \sum_{k\in\mathbb{N}} p_k = 1$) [Athreya and Ney, 1972, p. 7]. If the process is subcritical or critical ($m \leq 1$), then $q = 1$, i.e. the process dies out with certainty. If the process is supercritical ($m > 1$), then $q < 1$. Even then the process may die out if $p_0 > 0$. In our example, $f(s) = 0.2 + 0.8s^2 = s$ has the smallest non-negative solution $q = 0.25$. Although the cell population grows on average geometrically with a growth

rate of 0.6 per generation, the probability that the population, starting with one cell, will ever die out is therefore still a respectable $q = 0.25$. Because cells live, divide and die independently of each other, the probability that the population dies out once it has reached a size of n cells is $q_n = 0.25^n$. The risk of dying out is therefore largest for small populations and diminishes with increasing population size, i.e. populations that die out usually do so early. This example also illustrates another property of branching processes: they either die out or grow geometrically (exponentially for continuous-time branching processes described below). Branching processes are therefore very useful to study the early phase of a population growth or infection process, but they are not suitable to study the long-term dynamics of such a process.

Single-type Galton-Watson processes, i.e. single-type, discrete-time branching processes, can be generalised in two directions: towards multi-type and towards continuous-time branching processes (or a combination of both). I will now briefly present these generalisations.

In a multi-type Galton-Watson process, time is still discrete, but several types of individuals may exist (e.g. cells with or without a specific mutation) [Haccou et al., 2005, p. 21ff]. After one generation, each parent of type i may generate different numbers k_j of offspring of type j. Assume that there exist d types and that m_{ij} is the mean number of offspring of type j that a parent of type i has. The matrix $M = (m_{ij})_{1 \leq i,j \leq d}$ is then called the mean matrix and can be used to examine the average dynamics of the branching process. Under certain conditions (for example indecomposability of the branching process, so that each parent type can have offspring of any other type, - albeit perhaps only indirectly, over several generations), the matrix M has a positive largest eigenvalue ρ (called Perron root) with corresponding left eigenvector $\mathbf{u}$ and right eigenvector $\mathbf{v}$, i.e. $\mathbf{u} \cdot M = \rho \mathbf{u}$ and $M \cdot \mathbf{v} = \rho \mathbf{v}$. The Perron root ρ then has a similar role in the multi-type Galton-Watson process as the mean number m of offspring had in the single-type Galton-Watson process: if $\rho < 1$, $\rho = 1$ or $\rho > 1$, the multi-type Galton-Watson process is called subcritical, critical or supercritical, respectively. The process will die out with certainty if it is subcritical or critical. If the process is supercritical, $\rho > 1$, it has an extinction probability of less than one (but not necessarily zero), and if it does not become extinct, it will grow geometrically

with growth rate $\rho - 1$ per generation. If the left eigenvector $\mathbf{u}$ is scaled so that the sum of its components adds up to one, $\sum_{j=1}^{d} u_j = 1$, then the fraction of individuals of type j will tend towards u_j over time, i.e. the multi-type Galton-Watson process will tend towards a stable distribution of types, independent of the type with which the process started [Haccou et al., 2005, p. 28].

One can also drop the restriction of discrete time for branching processes. In a single-type, continuous-time branching process, there are no fixed generation times. Instead, the life spans of all individuals have identical distributions but are independent from each other. At the end of their life span, all individuals have k offspring ($k \in \mathbb{N}$) with the same probability distribution $\{p_k\}_{k\in\mathbb{N}}$, and again the number of offspring is independent of other individuals. The same is true for all their descendants. If the life span has an exponential distribution, the process is also a Markov process, called a Markov branching process, and is much easier to analyze than if the life span has a different distribution [Haccou et al., 2005, p. 59ff]. Consider a Markov branching process with density function of the form $\lambda e^{-\lambda t}$ for the life span $t \geq 0$. If the process starts with one individual, the expected (mean) number of individuals at time t is $x(t) = e^{\lambda(m-1)t}$, where m is the mean number of offspring at the end of the life span. Subcritical ($m < 1$) and critical ($m = 1$) processes will again die out with certainty, and supercritical processes ($m > 1$) have a positive probability to persist and will then grow exponentially.

Finally, multi-type, continuous-time branching processes allow for several types of individuals whose life-span distribution and whose distribution of offspring of different types may depend on the parent's type. I use such a model in chapter 2 to analyse the early phase of an IS infection. In my model, the type of the branching process corresponds to the number of IS copies in a host genome, and time is assumed to be continuous. I will explain details of the model in chapter 2.

1.4.2 Tau-leaping algorithm

Originally, the tau-leaping algorithm was developed to simulate the dynamics of chemically reacting systems [Gillespie, 2001], as an extension of the exact but computationally expensive Doob-Gillespie algorithm [Gillespie, 1977]. Several authors have subsequently used the

tau-leaping algorithm to simulate population or infection dynamics [Khalili and Armaou, 2008, Vaughan et al., 2012]. In chapter 4, I use the tau-leaping algorithm to simulate the infection dynamics inside a well-mixed subpopulation. The formulation of the algorithm below will therefore be adapted to accommodate this situation.

The basic algorithm works as follows [Gillespie, 2001]. Assume that a population consists of different types of individuals (e.g. uninfected, detrimentally infected, or beneficially infected by an IS). Everything that can happen to an individual is called an event, and there exist different classes of events. Each event class indicates a specific change of type, or the death or the propagation of an individual of a specific type. Two examples of event classes in our model are the detrimental infection of an uninfected cell (change of type) and the death of a beneficially infected cell. The rates with which those different event classes happen may depend on the numbers of individuals of any type. For example, cell death or IS infection by HGT may be density-dependent. For an event of any class, we can determine how it will affect the numbers of individuals of different types. For example, a detrimental infection of an uninfected cell will decrease the number of uninfected cells by one and increase the number of detrimentally infected cells by one. Suppose that we know the numbers of individuals of all different types at a specific time t_0, for example at the start of the simulation. We want to calculate the number of individuals after a time step of length τ (from which the algorithm got its name). Instead of assuming a fixed length τ for the time step, though, the idea in tau-leaping is to calculate the maximal length τ so that the rates of all event classes will approximately stay constant during one time step of length τ. In brief, we achieve this by choosing τ just as small as is necessary to ascertain that the prospective, relative changes in the rates of each event class, caused by prospective changes in the numbers of individuals of all types during one time step of length τ, will stay below a given threshold. Having calculated the length τ of the time step, during which we consider all rates to remain constant, we then determine for each event class the number of events occurring during τ by drawing a random number from a Poisson distribution with a mean equal to the product of τ and the event rate in that class at time t_0. Finally, we use all these events of all classes to calculate the numbers of individuals of all types at time $t_0 + \tau$. We then have executed one tau-leap.

The basic algorithm described in the last paragraph works well to simulate the population dynamics as long as none of the numbers of individuals of any type are small. If at any time the number of individuals of a type becomes small, there is a risk that during the next tau-leap the events that have been determined to happen would by chance require to decrease the number of individuals of that type to negative values. This is clearly unsatisfactory. To avoid such situations, we distinguish between noncritical and critical event classes at time t_0 [Cao et al., 2006]. Noncritical are those classes which will presumably not reduce any numbers of individuals that are involved in events of those classes to negative values during the next time step. For those event classes, we calculate τ_{noncrit} using the basic algorithm from the paragraph above. We call an event class critical, if during the next tau-leap, events of that class might result in negative numbers of individuals for at least one type. To avoid this, we determine the time span τ_{crit} until the next event of any of the critical event classes will happen by drawing a random number from an exponential distribution with mean $1/r$ where r is the sum of the rates of all critical event classes. The idea is to make sure that at most one event of any critical class happens during the next tau-leap. We therefore choose $\tau = \min(\tau_{\mathrm{crit}}, \tau_{\mathrm{noncrit}})$ as the length of the next τ-leap. The numbers of events of noncritical classes that occur during the next tau-leap are determined using the basic algorithm from the paragraph above. If $\tau = \tau_{\mathrm{noncrit}}$, no event of any critical class will happen during the tau-leap, and if $\tau = \tau_{\mathrm{crit}}$, exactly one event of any critical class will happen. We determine the class to which this event belongs by drawing a random number from a distribution with probabilities proportional to the rates of all critical event classes. Finally, we use all events of all classes (noncritical and perhaps critical) to calculate the numbers of individuals of all types at time $t_0 + \tau$. We then have again executed one tau-leap.

I use the tau-leaping algorithm described in the last paragraph to simulate the IS infection dynamics in a well-mixed subpopulation of 10^9 cells (see chapter 4).

1.5 Thesis outline

In my thesis, I studied different aspects of the infection dynamics of insertion sequences, which I now briefly summarize.

In chapter 2 [Bichsel et al., 2010], I investigate the early phase of an IS infection with a slightly detrimental IS, using a branching process model. This early phase is of interest because the IS can then easily get lost through chance events. I find that an IS infection starting with a single, infected cell has a high probability of dying out, and the median time to extinction is short. If an IS infection persists, its spread is very slow, and because of stronger detrimental effects of a high IS load, most infected genomes contain only few IS copies, in accordance with the IS count distribution determined in 728 fully sequenced prokaryote genomes.

In chapter 3 [Bichsel et al., 2013], I investigate the long-term fate of an IS and its host population, once the IS has overcome the high extinction risk of the early infection phase. I model the dynamics of host cells with different IS counts in their genome using a system of ordinary differential equations. Based on this model and the count distribution of a specific IS family in 1128 fully sequenced prokaryote genomes, I determine maximum likelihood estimates for the fitness effect of an IS and for the HGT rate. I find that the ISs of this family are likely to be effectively neutral or only slightly detrimental, and that the estimated HGT rate is well within the range of HGT rates reported by other authors. I also show that with the estimated parameters, the time needed to reach the prevalence of infected cells observed in the wild is unrealistically long. Occasional beneficial effects may thus have accelerated the infection process.

In chapter 4 [corresponding paper submitted], I investigate the role of space in the infection dynamics of an IS that can have both detrimental and beneficial fitness effects on its host. Using metapopulations of spatially distributed subpopulations, I simulate the spreading of an IS infection that starts in a single subpopulation. I find that the spatial structure of the metapopulation and dispersal between subpopulations are not limiting factors in the spread of an IS infection, but that factors within subpopulations (e.g. fitness effects and HGT) strongly influence the infection dynamics. I also show that the fate of an IS infection in spatially structured metapopulations depends critically on the initially infected subpopulation.

2. The Early Phase of a Bacterial Insertion Sequence Infection

Manuel Bichsel, Andrew D. Barbour, Andreas Wagner; *Theoretical Population Biology*, 2010, 78(4):278–288

Abstract

Bacterial insertion sequences are the simplest form of autonomous mobile DNA. It is unknown whether they need to have beneficial effects to infect and persist in bacterial populations, or whether horizontal gene transfer suffices for their persistence. We address this question by using branching process models to investigate the critical, early phase of an insertion sequence infection. We find that the probability of a successful infection is low and depends linearly on the difference between the rate of horizontal gene transfer and the fitness cost of the insertion sequences. Our models show that the median time to extinction of an insertion sequence that dies out is very short, while the median time for a successful infection to reach a modest population size is very long. We conclude that horizontal gene transfer is strong enough to allow the persistence of insertion sequences, although infection is an erratic and slow process.

2.1 Introduction

Ever since its discovery in the 1940s by Barbara McClintock [McClintock, 1950], mobile DNA has fascinated researchers. Why does it exist, and how does it persist? Some authors claim that mobile DNA ultimately needs to have beneficial effects on the host cell to be able to persist in the long term [Blot, 1994, Shapiro, 1999, Schneider and Lenski, 2004]. Other

authors disagree and think that mobile DNA is selfish DNA, which merely persists by replicating inside a host cell's genome and by infecting new hosts through sexual reproduction or horizontal gene transfer [Dawkins, 1976, Doolittle and Sapienza, 1980, Orgel and Crick, 1980, Charlesworth et al., 1994, Nuzhdin, 1999]. While even purely detrimental mobile DNA can spread in a sexually reproducing eukaryote population [Charlesworth et al., 1994], the persistence of detrimental mobile DNA in an asexually reproducing, prokaryote population is more difficult to explain.

Besides raising theoretical issues, the existence and persistence of certain classes of mobile DNA is also of practical interest. Some prokaryotic transposons – mobile DNA elements that move inside their host genome through a cut-and-paste process – carry antibiotic resistance genes [Berg, 1989, Kleckner, 1989], genes encoding toxins [So and McCarthy, 1980], or genes with new metabolic functions [Top and Springael, 2003]. Thus, transposons on the one hand contribute to an important public health threat by spreading antibiotic resistance among pathogens. On the other hand, transposons are also very useful tools in genetic engineering.

Prokaryotic transposons consist of two groups: simple and composite transposons. Simple transposons encode the proteins needed for their mobility themselves. Composite transposons contain two flanking insertion sequences, another class of mobile DNA. The insertion sequences encode the protein needed for the composite transposon's mobility.

Bacterial insertion sequences (ISs) are short DNA segments with a length of between 700 and 2700 bp [Chandler and Mahillon, 2002]. An IS usually codes for only one protein, transposase, which excises it from its current position in the genome and inserts it at a new position, a process called conservative transposition. Occasionally, instead of being cut-and-pasted, an IS is copy-and-pasted through replicative transposition. Replicative transposition increases the IS count per genome; however, ISs are sometimes also excised, thus decreasing the IS count. ISs are probably the simplest form of autonomous mobile DNA, encoding for just enough functionality to move and spread on their own inside a host genome. Currently, all ISs have been classified into 20 families, based on differences in their internal organization (open reading frames), in their transposases, in the nucleotide sequence at their ends, and in the nucleotide sequences they leave behind in the genome

after being excised [Chandler and Mahillon, 2002, Mahillon et al., 2009]. Individual ISs are named ISn, where n is an integer (e.g. IS*1*, IS*2* and IS*630*).

ISs pose a threat to host cells for at least two reasons. First, ISs can disable genes by inserting themselves into them. Second, if more than one IS is present in a genome, ISs can lead to the rearrangement of the whole host genome through homologous recombination [Galas and Chandler, 1989, Kleckner, 1989, Schneider and Lenski, 2004]. Therefore, although ISs can occasionally cause beneficial mutations [Hall, 1999, Schneider and Lenski, 2004], their general effect on the host cell is probably detrimental, especially if the IS count per genome is high.

Why then do ISs persist? When an IS first enters an uninfected host cell population, it occurs in only one or a few genomes of the population. It can then spread by horizontal gene transfer (HGT) to the genomes of other cells. This early phase of an IS invasion is crucial for its long-term fate and has parallels in the fate of a rare, slightly detrimental allele in a large population [Ohta, 1974]. HGT is a necessary condition for the persistence of a detrimental IS. But is HGT enough to allow IS persistence, or are (albeit rare) beneficial effects of ISs needed? We address this question by modeling the early phase of an invasion of a slightly detrimental IS into an uninfected bacterial host cell population as a branching process. Specifically, we first use our branching process models to compute the survival probability of an IS infection, its time to extinction if it becomes extinct, and the time to reach a given population size threshold if the infection persists. Last, we use our multi-type branching process model to derive the distribution of the IS count per infected cell genome, and we compare this distribution with the real IS count distribution in 728 fully sequenced bacterial genomes.

2.2 Models

In our models, we assume a large bacterial host cell population living at carrying capacity. Into this population, we introduce one cell infected with a single IS. We use a continuous-time, multi-type Markov branching process model to compute the IS survival probability, the time needed to reach a given population size threshold if the IS persists, and the IS count distribution [Haccou et al., 2005, Athreya and Ney, 1972]. We use a related birth-

and-death process model to analyse the time to extinction if the IS becomes extinct. Being stochastic processes, branching processes are particularly well-suited to model the early phase of an IS infection, given that the number of infected cells is still low and prone to strong random fluctuation. The use of branching process models in population genetics dates back to Fisher [Fisher, 1922] and Haldane [Haldane, 1927]. For introductions to branching processes and their use in biology, see [Athreya and Ney, 1972, Sewastjanow, 1975, Jagers, 1975, Kimmel and Axelrod, 2002, Haccou et al., 2005].

As we only model the early phase of an IS infection, we assume that the number of infected cells is always several orders of magnitude lower than the number of uninfected cells. We furthermore assume the cells to live in a well-mixed bulk environment, e.g. in seawater. In such an environment, each infected cell is surrounded by uninfected cells only and not influenced by any other infected cells, i.e. there is no HGT between infected cells.

We do not allow for immigration or emigration of cells, and as the host cell population lives at carrying capacity, the cell division rate b equals the base death rate d. For convenience, we choose $b = d = 1$ per cell generation. This choice of the cell division rate leads to the generation time being one time unit. As ISs are relatively short compared to their host genome (2.7 kbp at the most, versus e.g. around 4500 to 5500 kbp for the *E. coli* genome [Bergthorsson and Ochman, 1998]), we neglect the small additional cost needed in replicating ISs during cell division and assume the same birth rate b for infected cells as for uninfected cells. Empirical data suggest a death rate of infected cells with at most a linear dependence on the IS count per genome [Sawyer et al., 1987]. We assume a linearly increasing death rate of the form $d+js$ for infected cells, where j is the IS count per genome, and $s \ll d$ is the fitness cost per IS.

We allow for five event types that change the total IS count in the population: division of an infected cell, death of an infected cell, replicative transposition of an IS, excision of an IS, and HGT. In HGT, an IS is copied from an infected cell to an uninfected cell.

2.2.1 Multi-type model

Our multi-type model is inspired by and similar to the models used by Moody [Moody, 1988] and by Basten and Moody [Basten and Moody, 1991], but our model differs in the effect

of a cell's IS count on the cell's fitness, and, more importantly, instead of assuming a fixed bacterial generation time, we assume a continuous, exponentially distributed generation time. Although not strictly correct [Powell, 1955], an exponentially distributed generation time has been chosen to simplify calculations, because the branching process is then also a Markov process. In any case, our results will still be qualitatively correct if a better suited non-exponentially distributed generation time is assumed.

Some ISs down-regulate their transposition rate with increasing IS count per genome [Sawyer et al., 1987, Chandler and Mahillon, 2002]. An example is IS*10*, where the IS produces both a locally operating transposase and a globally operating negative regulator of transposase gene expression, so that with increasing IS count the transposase density at an IS site stays constant, while the density of the negative regulator increases. We include this effect in our model and assume the replicative transposition rate u per infected cell and per generation to be constant and independent of the cell genome's IS count (but see subsection 2.4.5 for a discussion of the effects of a nonconstant transposition rate). Furthermore, we assume excision events to be independent of each other. In our multi-type model, we therefore adopt a rate je of IS excision events per infected cell and generation, proportional to the genome's IS count j and the excision rate e per IS, where $e < u$ [Egner and Berg, 1981]. It is not known whether the IS count of a cell's genome influences the cell's HGT rate. But it is known that HGT is tightly regulated and depends on many internal and external factors [Dröge et al., 1999], of which the IS count of the donor cell is probably only a minor one. For simplicity, we assume a constant rate h of HGT per infected cell and per generation, independent of the cell genome's IS count (see subsection 2.4.5 for a discussion of the effects of a nonconstant HGT rate).

To avoid having to deal with an infinite-dimensional system, we assume an upper limit of $l = 50$ ISs per genome, except where noted otherwise. This is not a serious restriction, because only a very small proportion of infected cells in the wild has such a high IS count, and most infected cells harbor only a few ISs in their genome, as has already been seen before [Sawyer et al., 1987, Wagner, 2006, Touchon and Rocha, 2007], and as we also show in subsection 2.3.4.

Figure 2.1 shows the structure of the multi-type model, as defined by our assumptions.

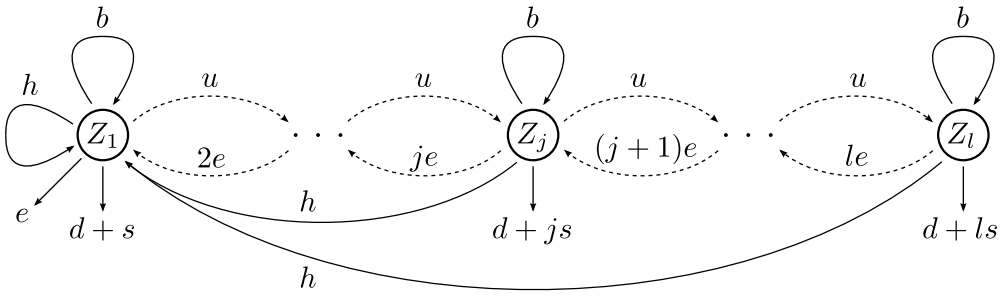

Figure 2.1: *Multi-type model design.* Z_k = number of cells with k ISs ($k \in \{1, \ldots, l\}$), b = birth rate per cell, d = base death rate per cell, u = replicative IS transposition rate per cell, e = IS excision rate per IS, h = HGT rate per cell, s = fitness cost per IS, and l = maximal IS count per genome (all rates are per host cell generation). Solid arrows indicate a change of total IS count and total infected cell count. Dashed arrows indicate a change of total IS count only.

A cell genome's IS count k, $k \in \{1, \ldots, l\}$, determines the cell's event rate a_k, i.e. the rate at which either a cell death, a cell birth, a replicative transposition event, an excision event, or an HGT event happen in a cell harboring k ISs:

$$a_1 = b + d + s + u + e + h$$

$$a_j = b + d + js + u + je + h \quad (1 < j < l)$$

$$a_l = b + d + ls + le + h,$$

where b and d are the birth and base death rates, respectively, s is the fitness cost per IS copy, u is the replicative transposition rate, e is the IS excision rate, and h is the rate of HGT.

The waiting time to the cell's next event is assumed to have an exponential distribution with mean $1/a_k$, and at the time of an event, the probabilities p_k of the five different event types are given by

IS count	cell div.	cell death	transp.	excision	HGT
1	$\frac{b+h}{a_1}$	$\frac{d+s+e}{a_1}$	$\frac{u}{a_1}$	0	0
$1 < j < l$	$\frac{b}{a_j}$	$\frac{d+js}{a_j}$	$\frac{u}{a_j}$	$\frac{je}{a_j}$	$\frac{h}{a_j}$
l	$\frac{b}{a_l}$	$\frac{d+ls}{a_l}$	0	$\frac{le}{a_l}$	$\frac{h}{a_l}$.

For a cell infected with one IS, excision is counted as cell death (uninfected cells are not

included in the model), and HGT is counted as cell division.

The event probabilities p_k are then used to define the vector-valued probability generating function $\mathbf{g}(\mathbf{z}) = \sum_{\mathbf{j}} \mathbf{p}(\mathbf{j})\mathbf{z}^{\mathbf{j}}$, $\mathbf{z} = (z_1, \ldots, z_l)$ (see 2.5.1). From the probability generating function, we derive the infinitesimal generating functions $\tilde{g}_k(\mathbf{z}) = a_k(g_k(\mathbf{z}) - z_k)$, and the infinitesimal generator A, which is defined as $A = (a_{ij}) = a_i b_{ij}$, where $b_{ij} = \frac{\partial g_i(\mathbf{z})}{\partial z_j}\Big|_{\mathbf{z}=\mathbf{1}} - \delta_{ij}$ [Athreya and Ney, 1972, p. 183 and 200], also shown in 2.5.1. The eigenvalue λ_0 of A with the largest real part is itself real. If λ_0 is negative, 0, or positive, the branching process is called subcritical, critical, or supercritical, respectively. If the branching process is subcritical or critical, it will become extinct with certainty; if the branching process is supercritical, it has a positive probability smaller than one of survival.

If the branching process is supercritical, there exist positive right and left eigenvectors $\mathbf{u} = (u_1, \ldots, u_l)$ and $\mathbf{v} = (v_1, \ldots, v_l)$ of the infinitesimal generator A, which can be scaled so that $\sum_{k=1}^{l} u_k = 1$ and $\sum_{k=1}^{l} u_k v_k = 1$. In the following, it will always be assumed that this scaling has been done for $\mathbf{u}$ and $\mathbf{v}$.

2.2.2 Single-type model

For the birth-and-death process model, we simplify the multi-type model by assuming that transposition and excision can be neglected, so that there is only one type of infected cell, bearing exactly one IS. The process state of the birth-and-death process model corresponds to the number of infected cells, and process state 0 is considered to be absorbing, meaning that the population of infected cells has become extinct. The birth and death rates per infected cell are $b + h$ and $d + s$, respectively, where again b and d are the birth and base death rates of a cell, h is the HGT rate, and s is the fitness cost of an IS.

Feller was the first to investigate this birth-and-death process [Feller, 1939]. Kendall derived the probability $P_n(t)$ of the process being in state n at time t [Kendall, 1948]. In our case, this probability is

$$P_n(t) = \begin{cases} \xi_t & \text{if } n = 0 \\ (1 - \xi_t)(1 - \eta_t)\,\eta_t^{n-1} & \text{if } n > 0 \end{cases} \tag{2.1}$$

where

$$\xi_t = \frac{(d+s)\left(1-e^{-(b+h-(d+s))t}\right)}{b+h-(d+s)\,e^{-(b+h-(d+s))t}} \text{ and } \eta_t = \frac{(b+h)\left(1-e^{-(b+h-(d+s))t}\right)}{b+h-(d+s)\,e^{-(b+h-(d+s))t}}.$$

At all times t, the state of the birth-and-death process is therefore zero (i.e. the process has become extinct) with probability ξ_t, and otherwise the state has a geometric distribution with parameter η_t.

2.2.3 Model parameters

We now turn to the parameters that we are using to analyze the models. Reliable rates for replicative transposition, IS excision and HGT are difficult to establish. However, in some cases at least their order of magnitude is known or can be estimated. Conservative transposition occurs with a rate of around 10^{-7} to 10^{-4} events per cell and host cell generation [Chandler and Mahillon, 2002,Kleckner, 1989]. We assume the replicative transposition rate to be a few orders of magnitude smaller [Tavakoli and Derbyshire, 2001]. IS excision rates are lower than replicative transposition rates [Egner and Berg, 1981]. For example, IS*10* is excised from the genome at a rate of around 10^{-10} per cell and host cell generation, whereas its conservative transposition rate is 10^{-4} per cell and host cell generation [Kleckner, 1989]. Similarly, transposon Tn*5*, a mobile DNA sequence flanked by two copies of IS*50*, has an excision rate of 10^{-6} to 10^{-5} and a conservative transposition rate of 10^{-3} to 10^{-2} [Berg, 1977]. HGT rates vary widely and depend on many environmental factors. For viral transduction in marine bacteria, rates of between $1.6\cdot10^{-8}$ and $3.7\cdot10^{-8}$ transductants per colony-forming unit have been reported [Jiang and Paul, 1998]. For the conjugational transfer of plasmids in diverse seawater bacteria, $2.3\cdot10^{-6}$ to $5.6\cdot10^{-5}$ transconjugants per recipient cell have been found after 3 days of incubation [Dahlberg et al., 1998]. For transformation involving epilithic bacteria from a river, *in* situ rates of $2.2\cdot10^{-6}$ to $1.0\cdot10^{-3}$ events per recipient cell have been reported per 24 hours incubation time [Williams et al., 1996]. Note that in this case, the transformation occurred in cells that were fixed on a surface, i.e. not in a well-mixed environment as we assume in our models. No information is available about the fitness cost caused by ISs. In our models, we therefore vary this cost

over a broad range of values.

Table 2.1 shows a summary of reported rates and rates used in our models.

Event	Reported rates		Model rates	
Transposition	Conservative	10^{-7} – 10^{-4}	Replicative	10^{-9} – 10^{-6}
Excision		10^{-10}		10^{-12} – 10^{-9}
HGT	Transduction	10^{-8}		
	Conjugation	10^{-6} – 10^{-5}	Total	10^{-7} – 10^{-4}
	Transformation	10^{-6} – 10^{-3}		
Fitness cost		–		10^{-12} – 10^{-6}

Table 2.1: Event rates reported by different authors (rates are converted into events per cell or IS and day), and corresponding parameter ranges used in our models. Model rates are per cell and cell generation or, in the case of the fitness cost, per IS and cell generation. Origin of reported rates: conservative transposition [Kleckner, 1989,Chandler and Mahillon, 2002], excision [Kleckner, 1989], transduction [Jiang and Paul, 1998], conjugation [Dahlberg et al., 1998], transformation [Williams et al., 1996].

2.2.4 Software

We use Mathematica version 7.0 to carry out the numerical and analytical model computations. With the exception of figure 2.7, we also use Mathematica to generate the figures in the Results section. Figure 2.7 has been generated by first counting ISs in fully sequenced bacterial genomes using IScan [Wagner et al., 2007], and then computing the IS count distribution using R, version 2.6.2 [R Development Core Team, 2008].

2.3 Results

2.3.1 The survival probability of an IS infection is small

The survival probability p_{surv} of an IS infection starting with one cell that is infected with a single IS is given by $p_{\text{surv}} = 1 - p_{\text{ext}}$, where p_{ext} is the infection's extinction probability. The extinction probability of an IS infection starting with one cell that is infected with k ISs is the k-th component of the smallest root $\mathbf{q} = (q_1, \ldots, q_l)$ of the infinitesimal generating function $\tilde{\mathbf{g}}(\mathbf{z})$ in the interval $[\mathbf{0}, \mathbf{1}]$ [Athreya and Ney, 1972, p. 205]. The survival probability of an infection starting with one cell that contains one IS in its genome can therefore be computed as $p_{\text{surv}} = 1 - q_1$.

Figure 2.2 shows the survival probability as a function of the relative difference between the HGT rate and the fitness cost of an IS, based on a numerical computation of q_1 for different parameter combinations.

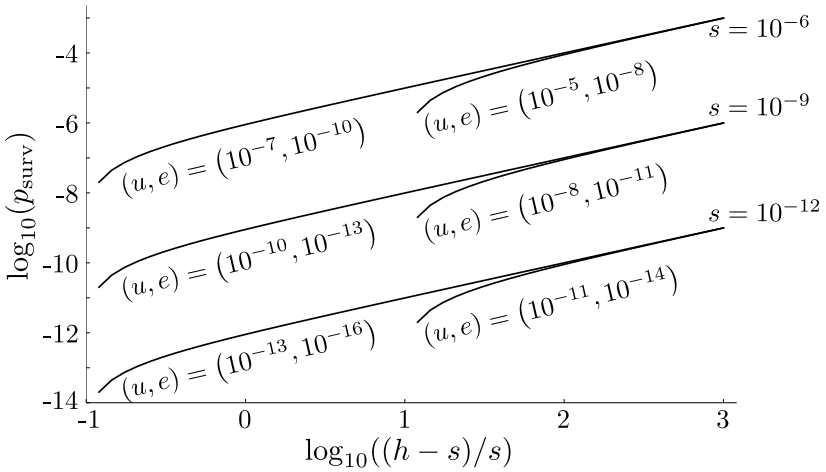

Figure 2.2: Computed survival probability p_{surv} of a population of infected cells, starting with one cell infected with a single IS, as a function of the relative difference $(h - s)/s$ between the HGT rate h and the fitness cost s, for different parameter combinations. Note the logarithmic scales. Parameter values: $b = d = 1$ gen.$^{-1}$, $(s, u, e) \in \{(10^{-12}, 10^{-13}, 10^{-16}), (10^{-12}, 10^{-11}, 10^{-14}), (10^{-9}, 10^{-10}, 10^{-13}), (10^{-9}, 10^{-8}, 10^{-11}), (10^{-6}, 10^{-7}, 10^{-10}), (10^{-6}, 10^{-5}, 10^{-8})\}$ gen.$^{-1}$, $l = 50$.

Figure 2.2 shows that $p_{\text{surv}} \approx h - s$, i.e. that the survival probability of an IS infection starting with one cell that is infected with one IS is approximately equal to the difference between the HGT rate and the fitness cost, at least if the replicative transposition rate u is smaller than the fitness cost s per IS. Only if $u > s$ does the infection's survival probability drop well below $h - s$ for low HGT rates h. The comparatively small excision rate does not have a significant effect on the infection's survival probability.

This result can be interpreted as follows: an IS infection can only persist if HGT is strong enough to overcome the mean fitness cost induced by ISs in infected cells (cf. figure 1). For replicative transposition rates that are lower than the fitness cost per IS, most cells will have only one IS. In that case, the survival probability of an infection will linearly depend on the difference $h - s$ between the HGT rate and the fitness cost induced by one IS. If, on the other hand, the replicative transposition rate is much larger than the fitness cost per IS, the population of infected cells includes many cells with higher IS counts, thus increasing the mean fitness cost per infected cell. This leads to a survival probability lower

than $h - s$.

The negative effect that a high replicative transposition rate has on the survival probability of an IS infection can also be demonstrated by computing the HGT rate h_{crit} at which the multi-type branching process is critical and will only just become extinct with certainty. h_{crit} can be computed by observing that λ_0, the eigenvalue with the largest real part of the infinitesimal generator A (see 2.5.1), must then be 0. Therefore, the constant term in the characteristic polynomial of A, which equals the determinant of A, must vanish. As h occurs only in the first column of A, the constant term linearly depends on h, and looking for its root, we find h_{crit}. Figure 2.3 shows h_{crit} as a function of s.

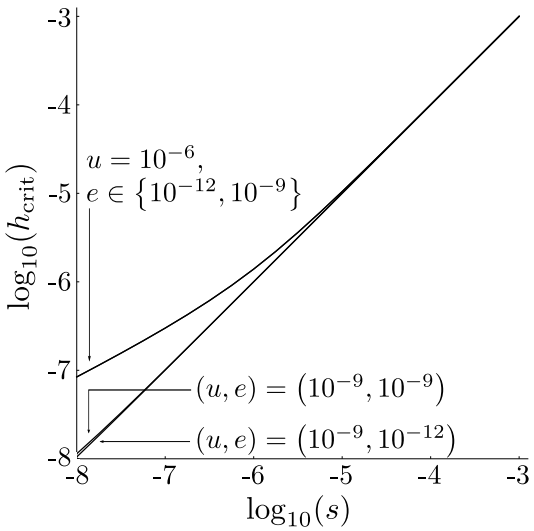

Figure 2.3: Computed critical HGT rate h_{crit} as a function of the fitness cost s, for different parameter combinations. Note the logarithmic scales. Parameter values: $b = d = 1$ gen.$^{-1}$, $(u, e) \in \{(10^{-9}, 10^{-12}), (10^{-9}, 10^{-9}), (10^{-6}, 10^{-12}), (10^{-6}, 10^{-9})\}$ gen.$^{-1}$, $l = 50$.

Figure 2.3 shows that for a fitness cost much larger than the replicative transposition rate u (infected cells then carry only one IS), the critical HGT rate is equal to the fitness cost. Figure 2.3 also shows that for a fitness cost coming near or falling below the replicative transposition rate (infected cells then carry on average more than one IS), the critical HGT rate is higher than the fitness cost per IS, because HGT has to compensate for a larger total fitness cost caused by a higher IS count per cell.

We will see later that the IS count distribution in infected cells is indeed strongly L-shaped, i.e. most infected cells contain only one or at most a few ISs in their genome (see subsection 2.3.4). We can therefore use the birth-and-death process model as an approx-

imation to our multi-type branching process model. In this single-type model, we can analytically confirm that $p_{\text{surv}} \approx h - s$ for small values of h and s. To do this, we observe that our birth-and-death process only survives if it does not get absorbed in state 0. Using (2.1), we therefore get

$$p_{\text{surv}} = 1 - \lim_{t\to\infty} P_0(t) = 1 - \frac{d+s}{b+h}.$$

Remembering that $b = d = 1$ gen^{-1}, and linearizing around $h = s = 0$ gen^{-1} then gives

$$p_{\text{surv}} \approx h - s.$$

Haldane [Haldane, 1927], following an idea of Fisher [Fisher, 1922], showed that a dominant mutant gene with a small selective advantage s, so that the expected number of offspring is $1 + s$, has a probability of about $2s$ to persist in a random mating population. Observe that in our case, the selective advantage of a cell that harbors an IS is $(h - s)/2$, as the cell's expected number of offspring is $2 \cdot (b + h)/(b + d + h + s) \approx 1 + (h - s)/2$ for $b = d = 1$ gen^{-1} and small h and s.

2.3.2 The time to extinction of an IS infection is short

According to the last subsection, the vast majority of IS infections die out. Again considering that IS infections are dominated by cells with only a few ISs (see subsection 2.3.4), we use the single-type birth-and-death process model to compute the time to extinction of an IS infection that becomes extinct. We start with one infected cell in an uninfected host cell population. We then use the process state probability given in (2.1), observing that the probability of the birth-and-death process ever becoming extinct is given by $\lim_{t\to\infty} P_0(t)$. Therefore, using our assumption that $b = d = 1$ gen^{-1}, the cumulative distribution function F of the time to extinction T_0, conditioned on the branching process becoming extinct, is

$$F(t) = P(T_0 \le t | T_0 < \infty) = \frac{P_0(t)}{\lim_{t\to\infty} P_0(t)}.$$

As we have shown earlier, only in the case $h > s$ is there a positive probability of the birth-and-death process not becoming extinct. The distribution function is then

$$F(t) = \frac{(1+h)\left(1 - e^{-(h-s)t}\right)}{1 + h - (1+s)\, e^{-(h-s)t}},$$

and the corresponding probability density function of the time to extinction, conditioned on the branching process becoming extinct, is

$$f(t) = \frac{dF(t)}{dt} = \frac{(1+h)(h-s)^2\, e^{-(h-s)t}}{\left(1 + h - (1+s)\, e^{-(h-s)t}\right)^2}.$$

Figure 2.4 shows the density of T_0 for different parameter combinations of the fitness cost s and the HGT rate h, where always $h > s$.

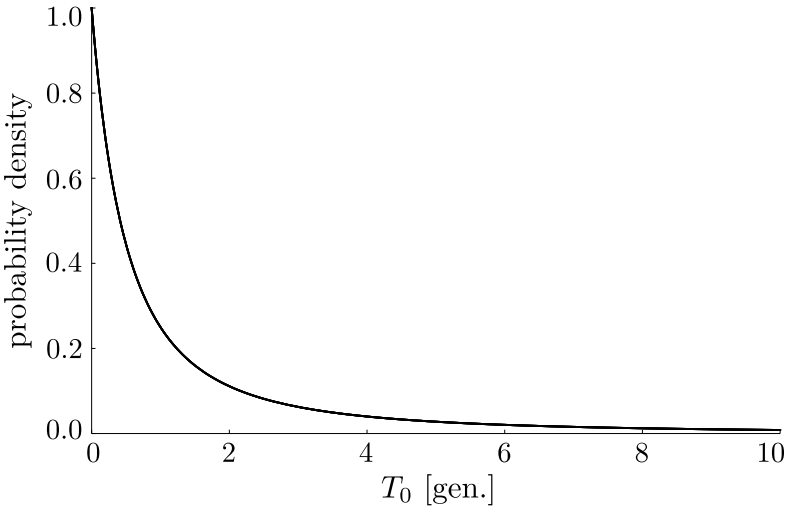

Figure 2.4: Probability density function of the time to extinction T_0, for different parameter combinations. Parameter values: $b = d = 1$ gen.$^{-1}$, $(s, h) \in \{(10^{-12}, 10^{-7}), (10^{-12}, 10^{-4}), (10^{-6}, 10^{-4})\}$ gen.$^{-1}$. The single line is an overlay of the graphs obtained when using the three parameter value combinations of s and h indicated above.

Figure 2.4 shows that first, the time to extinction is not strongly influenced by the fitness cost of an IS and by the HGT rate, and second, the distribution of the time to extinction is very skewed. Because of the latter observation, the median $T_{0,\text{med}}$ of the time to extinction is more useful to report than the mean. We use the distribution F of the time to extinction to obtain the median time to extinction. To this end, we first transform F algebraically

and then linearize the transformed expression around $h = s = 0$ gen^{-1}:

$$F(t) = \frac{1}{1 + \frac{\frac{h-s}{1+h} e^{-(h-s)t}}{1-e^{-(h-s)t}}} \approx \frac{t}{t+1} + \frac{1}{2}\left(\frac{t}{t+1}\right)^2 \left(\frac{t+2}{t}h - s\right)$$

Solving the equation $F(t) = 1/2$ for t and then again linearizing around $h = s = 0$ gen^{-1} gives the median time

$$T_{0,\text{med}} \approx \frac{\sqrt{1+h+h^2-s} - h}{1+h-s} \approx 1 - \frac{3h-s}{2}$$

The median time to extinction of an IS infection that becomes extinct therefore almost linearly depends on $3h - s$, but is dominated by the comparatively large constant 1. In this short time, replicative transposition and excision cannot take effect, which adds justification to our use of the birth-and-death process model.

2.3.3 The time an IS infection needs to attain a modest size threshold is long

Only a small fraction of IS infections survives. In a branching process, the surviving populations go into exponential growth after having lingered at lower population sizes during a random time period [Haccou et al., 2005, p. 158], [Athreya and Ney, 1972, p. 206], where they have been under strong threat of extinction. We first use our multi-type branching process model to numerically compute the time needed by a surviving population of infected cells to reach a given population size threshold. We then use our single-type birth-and-death process model to analytically confirm our numerical results from the multi-type model.

In a supercritical, irreducible, multi-type branching process with finite second moment as described by our multi-type model, the following holds [Sewastjanow, 1975, pp. 257–258]:

1. The random variable $W_k^m(t) := \frac{Z_k^m(t)}{v_k e^{\lambda_0 t}} \overset{t\to\infty}{\longrightarrow} W^m$ for any $m \in \{1, \ldots, l\}$ and $k \in \{1, \ldots, l\}$, where $Z_k^m(t)$ is the number of cells of type k at time t, starting with one cell of type m at time $t = 0$, and where v_k is the k-th component of the scaled left eigenvector $\mathbf{v}$ to the eigenvalue λ_0 of the infinitesimal generator A defined in 2.5.1.

2. The characteristic function $\varphi^m(x) = \mathbb{E}\left(e^{iW^m x}\right)$ of W^m, where $i = \sqrt{-1}$, obeys the

system of ordinary differential equations $\frac{d\varphi^m(x)}{dx} = \frac{\tilde{g}^m(\varphi^1(x),\ldots,\varphi^l(x))}{\lambda_0 x}$, with $\varphi^m(0) = 1$ for $m \in \{1,\ldots,l\}$, where $\tilde{g}^m$ is the infinitesimal generating function.

The ordinary differential equation system can be numerically solved for the characteristic functions $\varphi^m(x)$, $m \in \{1,\ldots,l\}$ (see 2.5.2 for details of the system). By the Fourier inversion theorem, the probability density f^1 of the random variable W^1 can be reconstructed from its characteristic function φ^1 as $f^1(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\varphi^1(x)dx$. From W^1, in turn, the number $Z_k^1(t)$ of infected cells with k ISs at time t (large enough) in a population that has been infected with one cell containing one IS in its genome can be derived as $Z_k^1(t) \approx v_k e^{\lambda_0 t} W^1$. The total size of the population of infected cells is then $Z(t) := \sum_{k=1}^{l} Z_k^1(t) \approx e^{\lambda_0 t} W^1 \sum_{k=1}^{l} v_k$. Therefore, the time T_N to the threshold N is

$$T_N = \frac{1}{\lambda_0}\left[\ln(N) - \ln(W^1) - \ln\left(\sum_{k=1}^{l} v_k\right)\right].$$

We again use the median to characterize the time to threshold and get

$$T_{N,\text{med}} = \frac{1}{\lambda_0}\left[\ln(N) - \ln(W^1_{\text{med}}) - \ln\left(\sum_{k=1}^{l} v_k\right)\right],$$

where W^1_{med} is the median of the random variable W^1, which can be computed using the density f^1 of W^1. Figure 2.5 shows the median time to a threshold of 10^8 infected cells versus the difference between the HGT rate and the fitness cost, for different fitness costs s, replicative transposition rates u and excision rates e. $N = 10^8$ is still a comparatively small threshold in a population of bacterial cells. In a bulk environment like seawater, for example, $8.4 \cdot 10^8$ to $2.5 \cdot 10^{10}$ bacterial cells per liter have been counted [Thompson et al., 2004]. And still, the threshold is large enough to guarantee a negligible extinction probability once it has been attained by the population of infected cells. Because of the computational complexity involved in calculating the time to threshold, the maximal number of ISs per cell genome had to be reduced from $l = 50$ to $l = 5$. This is not a strong limitation, since the population of infected cells is dominated by cells that harbor only one or very few ISs.

Figure 2.5 shows that the median time to threshold is approximately inversely proportional to $h - s$ for large thresholds N, e.g. $T_{N,\text{med}} = 55.5 \cdot (h-s)^{-0.82}$ for $s = 10^{-8}$ gen.$^{-1}$ and $N = 10^8$. We have confirmed that for larger thresholds the approximation to inverse

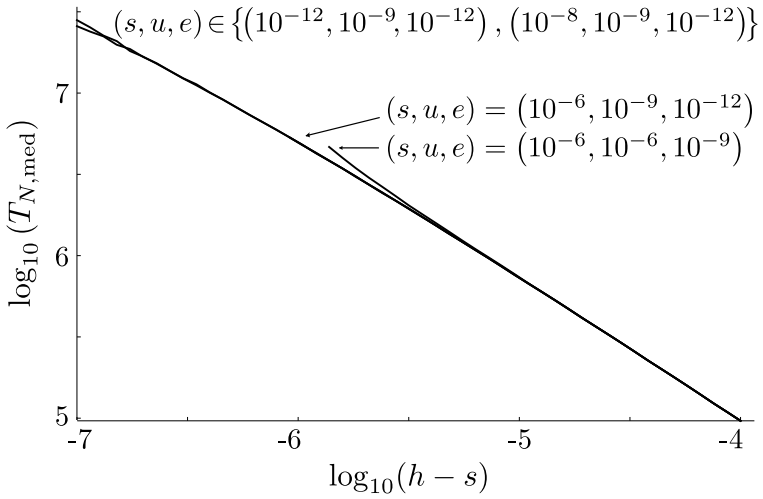

Figure 2.5: Computed median time $T_{N,\mathrm{med}}$ to a threshold of $N = 10^8$ infected cells as a function of the difference $h - s$ between the HGT rate and the fitness cost, for different parameter combinations. Note the logarithmic scales. Parameter values: $b = d = 1$ gen.$^{-1}$, $(s,u,e) \in \{(10^{-12},10^{-9},10^{-12}), (10^{-8},10^{-9},10^{-12}), (10^{-6},10^{-9},10^{-12}), (10^{-6},10^{-6},10^{-9})\}$ gen.$^{-1}$, $l = 5$. Because computing the characteristic function is feasible only for moderate fitness costs, not all graphs extend to the full range of the difference between the HGT rate and the fitness cost. The arrows mark the beginnings of the curves with the corresponding parameter sets.

proportionality becomes even better, e.g. $T_{N,\mathrm{med}} = 35.5 \cdot (h-s)^{-0.93}$ for $s = 10^{-8}$ gen.$^{-1}$ and $N = 10^{12}$ (graph not shown). This is because first, for large thresholds N, the population dynamics of the supercritical branching process is dominated by the exponential growth phase; second, the time spent in the exponential growth phase is inversely proportional to the growth rate, which is identical to the eigenvalue λ_0 of the infinitesimal generator A; third, at least for h much larger than s if $u \geq s$, λ_0 is approximately equal to the difference $h - s$ between the HGT rate and the fitness cost (see figure 2.6).

Figure 2.6 shows that if $s > u$ or $h > u$, $\lambda_0 \approx h - s$. Using a linear regression on the data shown in figure 2.6 that is restricted to $s = 10^{-8}$ gen.$^{-1}$ shows that $\lambda_0 \approx 1.00060 \cdot (h-s)^{1.00005}$. But λ_0 is smaller than $h - s$ (and can even drop below zero) if $u > s$ and $u \geq h$, because the population of infected cells is then no longer dominated by cells with only one IS, and HGT cannot replace fast enough the cells dying due to the increased total fitness cost per cell.

Because the population of infected cells is dominated by cells with only one IS, the single-type model is a good approximation to the multi-type model. We now use the

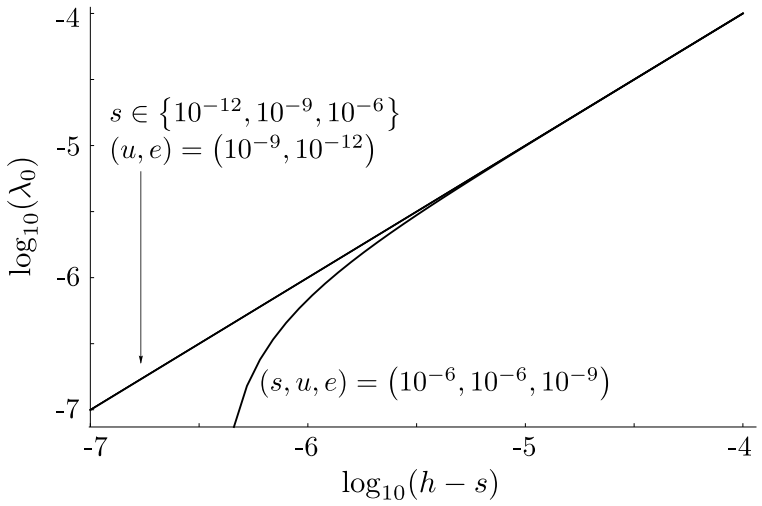

Figure 2.6: Growth rate λ_0 as a function of the difference $h - s$ between the HGT rate and the fitness cost, for different parameter combinations. Note the logarithmic scales. Parameter values: $b = d = 1$ gen.$^{-1}$, $(s, u, e) \in \{(10^{-12}, 10^{-9}, 10^{-12}), (10^{-8}, 10^{-9}, 10^{-12}), (10^{-6}, 10^{-9}, 10^{-12}), (10^{-6}, 10^{-6}, 10^{-9})\}$ gen.$^{-1}$, $l = 50$.

birth-and-death process model to analytically show that the median time to threshold is in fact approximately inversely proportional to the difference between the HGT rate and the fitness cost. Let again $Z(t)$ be the size of the population of infected cells at time t. Then $Z(t)/e^{(b+h-(d+s))t}$ is a nonnegative Martingale, and thus $\lim_{t\to\infty} Z(t)/e^{(b+h-(d+s))t} = W$ almost surely exists [Athreya and Ney, 1972, p. 111]. W is a random variable that is zero with probability $P(W = 0) = \frac{d+s}{b+h}$ and otherwise has an exponential distribution with rate parameter $\frac{b+h-(d+s)}{b+h}$ [Harris, 1951, p. 319].

From $\lim_{t\to\infty} Z(t)/e^{(b+h-(d+s))t} = W$, we get $\ln(Z(t)) - (b + h - (d + s))t \approx \ln(W)$ if t is large. Therefore, the time T_N to reach the threshold N, on the condition that it is reached, is $T_N \approx \frac{1}{b+h-(d+s)}[\ln(N) - \ln(W)]$. Using this approximation, we get for T_N the distribution function

$$
\begin{aligned}
P(T_N \le t) &= P\left(\frac{1}{b+h-(d+s)}(\ln(N) - \ln(W)) \le t\right) \\
&= 1 - P\left(W < N\, e^{-(b+h-(d+s))\,t}\right) \\
&= 1 - \int_0^{N\, e^{-(b+h-(d+s))\,t}} \frac{b+h-(d+s)}{b+h}\, e^{-\frac{b+h-(d+s)}{b+h}x}\, dx \\
&= \exp\left\{-\exp\left[-\left(x - \frac{\ln\left(N\frac{b+h-(d+s)}{b+h}\right)}{b+h-(d+s)}\right) \Big/ \frac{1}{b+h-(d+s)}\right]\right\}.
\end{aligned}
$$

This means that T_N has a Gumbel distribution, $P(T_N \leq t) = \exp\left(-e^{-(x-a)/b}\right)$ with parameters $a = \frac{1}{b+h-(d+s)} \ln\left(N\frac{b+h-(d+s)}{b+h}\right)$ and $b = \frac{1}{b+h-(d+s)}$, see [Johnson et al., 1995, p. 2], and therefore the median time to threshold is

$$\begin{aligned} T_{N,\text{med}} &= \frac{1}{b+h-(d+s)} \ln\left(N\frac{b+h-(d+s)}{b+h}\right) - \frac{1}{b+h-(d+s)} \ln(\ln(2)) \\ &= \frac{1}{b+h-(d+s)} \left[\ln\left(N\frac{b+h-(d+s)}{b+h}\right) - \ln(\ln(2))\right] \\ &\approx \frac{1}{h-s} \ln(N) \quad \text{if } N \text{ big and } b = d. \end{aligned}$$

This result shows that for large population size thresholds, the median time to threshold is approximately inversely proportional to the difference $h - s$ between the HGT rate and the fitness cost, and that the proportionality constant is the natural logarithm of the threshold size N.

2.3.4 The IS count distribution is biased towards low IS counts

The IS count distribution is the link between our model and real data. We demonstrate that our multi-type branching process model can adequately reproduce the real IS count distribution. Figure 2.7 shows the IS count distribution of the six most abundant ISs IS*1*A, IS*2*, IS*4*, IS*5*, IS*110* and IS*630*, which occur in at least 20 of the 728 bacterial genomes that have been fully sequenced as of June 2009. We obtained the necessary genome sequences from the National Center for Biotechnology Information, NCBI [NCBI, 2010], and we obtained the reference sequences of the ISs from the IS Finder database [Mahillon et al., 2009]. We used our previously published software IScan to identify and count ISs in the genomes, analogous to our earlier work [Wagner et al., 2007], but for a larger number of genomes.

Figure 2.7 shows that for each of the six most abundant ISs we examined, on average only 31 out of 728 sequenced bacterial genomes contain a minimum of one copy. The IS count distribution is L-shaped: most genomes contain none of these six ISs, a small number of genomes have up to a dozen copies of these ISs, and only a few genomes contain more than a dozen copies, although there are a few genomes containing many ISs. Among the six ISs we examined, only IS*1*A and IS*5* have more than 50 copies in some bacterial

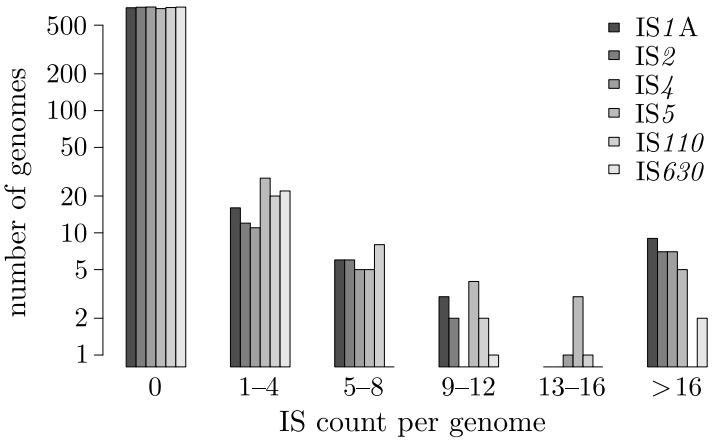

Figure 2.7: IS count distribution of the six most abundant ISs in 728 fully sequenced bacterial genomes (June 2009). Note the logarithmic vertical axis.

genomes: the seven sequenced *Shigella* genomes contain between 105 and 228 copies of IS*1*A, and *Xanthomonas oryzae* contains 53 copies of IS*5*. Of the 14 other, less abundant ISs we examined, only IS*481* and IS*982* have more than 50 copies in a genome: *Bordetella pertussis* contains 233 copies of IS*481* (all other genomes contain at most 11 copies of IS*481*), and *Lactotcoccus lactis cremoris* contains 56 copies of IS*982* (all other genomes contain at most 3 copies of IS*982*).

We do not distinguish between different prokaryotic species in the data of figure 2.7, because, especially for prokaryotes, HGT occurs across species boundaries [Gogarten and Townsend, 2005,Sørensen et al., 2005]. It is known that many ISs show DNA target specificities of varying degrees [Chandler and Mahillon, 2002]. For example, while IS*1* just prefers AT-rich regions, IS*4* is known to insert into DNA sequences of the form AAA–N_{15-20}–TTT [Zerbib et al., 1985, Mayaux et al., 1984]. In practice, target specificity is probably not strong enough to be a limiting factor in the IS count distribution.

We now derive the model's IS count distribution by pointing out that for our multi-type branching process, the limit $\lim_{t\to\infty} \frac{\mathbf{Z}(t)}{e^{\lambda_0 t}} = W\mathbf{v}$ almost surely exists, where $\mathbf{Z}(t) = (Z_1(t), \ldots, Z_l(t))$ is the vector of population sizes of infected cells with IS count $k \in \{1, \ldots, l\}$ at time t, W is a random variable (independent of the cell genome's IS count), and $\mathbf{v} = (v_1, \ldots, v_l)$ is the scaled left eigenvector to the eigenvalue λ_0 of the infinitesimal generator A [Athreya and Ney, 1972, p. 206]. Therefore, if $\mathbf{v}$ is rescaled so that $\sum_{k=1}^{l} v_k = 1$, its components $v_1, \ldots, v_l$ denote the limit distribution of IS counts in infected cells.

Figure 2.8 shows the computed limit distributions of IS counts per genome as a function of the HGT rate, for different parameter combinations. These limit distributions are approached asymptotically after the first IS infection occurred.

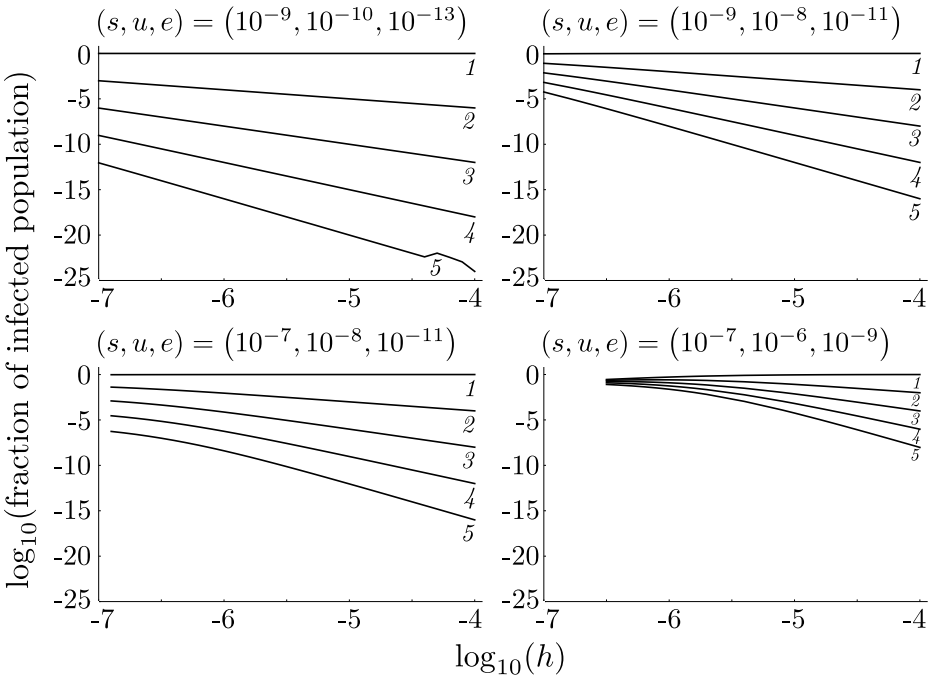

Figure 2.8: Computed IS count distribution as a function of the HGT rate h, for different parameter combinations. Note the logarithmic scales. Parameter values: $b = d = 1$ gen.$^{-1}$, $(s,u,e) \in \{(10^{-9}, 10^{-10}, 10^{-13}), (10^{-9}, 10^{-8}, 10^{-11}), (10^{-7}, 10^{-8}, 10^{-11}), (10^{-7}, 10^{-6}, 10^{-9})\}$ gen.$^{-1}$, $l = 50$ (but at most 5 ISs per infected cell are shown). The numbers in italics indicate the IS count per genome.

Figure 2.8 shows that for the broad parameter range used in our model, most infected cells contain only one IS. The decrease in the fraction of cells with two, three, or more ISs per genome gets even steeper for higher HGT rates. This result can be understood by noting that the IS count distribution in our multi-type model is determined by the replicative transposition rate u opposing the fitness cost s per IS and the HGT rate h (the excision rate e is too small to be of any importance). As $h > s$ is necessary for a persisting infection (see subsection 2.3.1), we can distinguish between three scenarios: $u > h > s$, $h > u > s$, and $h > s > u$. In the first scenario $u > h > s$, replicative transposition increases the IS count of cells faster than new cells can be infected with one IS. Therefore, the IS count distribution gets shifted towards higher values until an equilibrium is reached with the increasing total fitness cost per cell. In the second and third scenarios $h > u > s$

or $h > s > u$, HGT infects new cells faster with one IS than the IS count of already infected cells can increase. Therefore, the IS count distribution is strongly L-shaped. Considering our model parameter range, the latter two scenarios are more probable, and therefore, the IS count distribution in our model is generally L-shaped. Because $h > s$ is a necessary condition for IS infection persistence, no IS count distribution can be shown in figure 2.8 for $h < s$. In fact, an infection can become extinct with certainty even for h slightly larger than s if the IS count distribution is no longer strictly dominated by cells with one IS (see the lower right graph in figure 2.8).

2.4 Discussion

An IS that provides a sufficiently large benefit to its host can rapidly rise to fixation through natural selection [Hall, 1999, Schneider and Lenski, 2004]. We are interested in the more challenging scenario, where an IS is slightly detrimental. When newly introduced into an uninfected host cell population, such an IS faces a situation analogous to that of a slightly detrimental mutant allele that has newly emerged in a population. Its frequency in the population is subject to random drift, and it is easily driven to extinction [Moran, 1962, Ohta, 1974, Kimura, 1983]. However, this analogy with population genetics is limited: most population-genetic models are neither concerned with HGT, which can increase the number of cells carrying an IS for reasons different from selection and genetic drift, nor do they take into account the possibility of a genetic element increasing its number (and therefore its fitness cost) in a genome. Here, we focus on the interplay between HGT and other factors influencing the persistence of an infection with mobile genetic elements that can autonomously reproduce and increase their own number in an infected genome.

2.4.1 Survival probability

The linear dependency of p_{surv} on $h - s$ for low replicative transposition rate means that HGT stands in direct opposition to the selection against ISs. Specifically, an IS infection will only survive if the HGT rate h is higher than the fitness cost s of an IS. However, even if ISs have no fitness cost, the survival probability of an IS infection starting with one infected cell is small, because HGT rates are generally small. In a bulk environment

(e.g. seawater), HGT rates are probably at most 10^{-5} to 10^{-4} events per infected cell and generation [Dahlberg et al., 1998]. This range of the HGT rate provides an upper bound for the difference $h - s$. Even for neutral ISs, the survival probability of an IS infection starting with one infected cell would therefore be 10^{-4} at most.

2.4.2 Time to extinction

The median time to extinction $T_{0,\text{med}} \approx 1 - (3h - s)/2$ is dominated by the comparatively large constant 1. This means that half of the IS infections that die out do so in merely one generation. However, the distribution of the time to extinction is highly right-skewed. Some infections can therefore survive for a much longer time before they eventually die out.

The relationship $T_{0,\text{med}} \approx 1 - (3h - s)/2$ seems paradoxical at first, as the median time to extinction decreases with *increasing* HGT rate and/or *decreasing* fitness cost. However, this is due to the following bias: we are examining only infections that become extinct, and with increasing HGT rate and/or decreasing fitness cost, populations of infected cells tend to spend less time lingering at low population sizes before they either die out or begin to grow. In other words, an infection's fate is determined more quickly for increased HGT rate and/or decreased fitness cost, thereby reducing the median time to extinction.

2.4.3 Time to threshold

The time to threshold can be very long, especially if the HGT rate is only slightly higher than the fitness cost and therefore their difference almost vanishes. For the upper bound $h \in [10^{-5}, 10^{-4}]$ gen.$^{-1}$ of $h - s$ we used before, the median time to reach a population size threshold of 10^8 infected cells is between 10^5 and $10^{5.8} = 6.3 \cdot 10^5$ generations (see figure 2.5). Generation times of bacteria living in the wild vary broadly, but with an assumed generation time of one day for *E. coli* [Gibbons and Kapsimalis, 1967, Savageau, 1983], the median time to threshold for these large HGT rates is between 300 and 1700 years. As the time to threshold is right-skewed, it can sometimes be much longer. Because our information about IS infections stems from limited samples, such long times to threshold would in practice make it difficult to detect many IS infections, even if they were successful in the end.

2.4.4 IS count distribution

Within broad parameter ranges, our model predicts that a large majority of infected cells harbor only one IS per genome, and the fraction of cells with more than one IS drops quickly with increasing IS count. This holds even more for high HGT rates. The predicted distribution, biased towards very low IS counts, is corroborated by empirical data from more than 700 genomes, and it has also been observed in previous work based on a smaller number of genomes [Sawyer et al., 1987, Touchon and Rocha, 2007].

If the fitness cost is larger than the replicative transposition rate, the IS count distribution is highly skewed over the whole range of used HGT rates, with most cells harboring only one IS (see figure 2.8). To get an IS count distribution similar to the empirical distribution shown in figure 2.7, the fitness cost probably has to be somewhat smaller than the replicative transposition rate. The replicative transposition rate, in turn, is very low. We assume it to be in the interval $u \in [10^{-9}, 10^{-6}]$ gen.$^{-1}$. Our models therefore suggest that ISs might be effectively neutral in their effects on the host cell.

2.4.5 Effects of nonconstant HGT and transposition rates

In our model, we assume the replicative transposition rate and the HGT rate to be independent of the cell's IS count. We now discuss an alternative scenario, where the replicative transposition rate and/or the HGT rate linearly increase with the cell's IS count. Specifically, we discuss the effects of these scenarios on the IS count distribution, the survival probability of an IS infection, and the time to threshold. We do not discuss the effects on the extinction probability of an IS infection, because extinction happens fast and does not leave much time for transposition and HGT, and because our birth-and-death process model does not include transposition.

If the replicative transposition rate linearly increases with the IS count, the balance of forces determining the IS count distribution shifts: replicative transposition is strengthened in its opposition against fitness cost and HGT. Infected cells reach higher IS counts than if the replicative transposition rate is constant; although in most cases, the IS count distribution is still dominated by cells with one or a few ISs. Only if the replicative transposition rate is larger than the HGT rate (and therefore larger than the fitness cost), then the IS

count distribution is dominated by cells with the highest IS count allowed in the model. This is an unrealistic scenario and not consistent with the observed IS count distribution. A shift towards higher IS counts increases the fitness cost and therefore reduces the survival probability; although only slightly so, as long as the IS count distribution is still dominated by cells with one or only a few ISs. For the same reason, the time to threshold does not noticeably change (but remember that for the time to threshold, we have to restrict our model to a maximum of $l = 5$ ISs per cell).

If the HGT rate linearly increases with the IS count, the IS count distribution shifts towards lower values, as more cells get infected with one IS. Together, the higher infection rate and the lower fitness cost induced by only one IS increase the survival probability of an infection, especially for HGT rates only slightly larger than the fitness cost of an IS. A higher infection rate and a lower fitness cost also slightly decrease the time to threshold.

If both the replicative transposition rate and the HGT rate increase linearly with the IS count, two opposing forces in shaping the IS count distribution are strengthened: infected cells will reach higher IS counts, and at the same time, cells with higher IS counts will infect more cells with only one IS. The IS count distribution then shifts towards higher IS counts, but less so than when only the replicative transposition rate linearly increases. The survival probability, on the other hand, is similar to the one observed when only the HGT rate linearly increases: although cells with higher IS counts bear a higher fitness cost, they also infect more cells with an IS and keep the IS infection spreading. For this reason, the time to threshold is also slightly lower than with constant replicative transposition and HGT rate, albeit not as low as when only the HGT rate linearly increases with the IS count.

2.4.6 Caveats

We here discuss the limitations of our analysis, some of which are caused by our model assumptions, whereas others are caused by limited data.

First, in our branching process models, we assume a well-mixed environment, where infected cells are surrounded by uninfected cells and where they are not clustered. The models are therefore not valid for bacteria living in a spatially structured environment, e.g. in a biofilm. Second, we assume that an infection starts with one cell that is infected with one

IS. We note that in naturally occurring bacterial populations, the prevalence of infected cells is low (see [Wagner, 2006, Touchon and Rocha, 2007] and figure 2.7). Therefore, even if many new bacterial cells are introduced into an uninfected host cell population, probably only a few of these new cells are infected. This justifies our assumption. Third, we restrict HGT to transferring an IS copy only into uninfected cells. Again, this is no serious restriction: first, we only consider the early phase of an IS infection, with a low number of infected cells, and second, we assume infected cells to be well-mixed with and surrounded by uninfected cells, so that HGT into already infected cells can be neglected.

2.5 Appendix

2.5.1 Models: Multi-type model

The probability generating function of a multi-type branching process is defined as

$$\begin{aligned}\mathbf{g}(\mathbf{z}) &= \sum_{\mathbf{j}} \mathbf{p}(\mathbf{j})\mathbf{z}^{\mathbf{j}} \\ &= \left(\sum_{(j_{11},\ldots,j_{1l})} p_1(j_{11},\ldots,j_{1l}) z_1^{j_{11}} \cdot \ldots \cdot z_l^{j_{1l}}, \ldots, \sum_{(j_{l1},\ldots,j_{ll})} p_l(j_{l1},\ldots,j_{ll}) z_1^{j_{l1}} \cdot \ldots \cdot z_l^{j_{ll}} \right),\end{aligned}$$

where $p_k(j_{k1},\ldots,j_{kl})$ is the probability of a particle of type k (here: a cell with k ISs) to produce $j_{k1},\ldots,j_{kl}$ particles of type $1,\ldots,l$. In our case, we get the following probability generating function:

$$\begin{aligned}g_1(\mathbf{z}) &= \frac{b+h}{a_1} z_1^2 + \frac{d+s+e}{a_1} + \frac{u}{a_1} z_2 \\ g_j(\mathbf{z}) &= \frac{b}{a_j} z_j^2 + \frac{d+js}{a_j} + \frac{u}{a_j} z_{j+1} + \frac{je}{a_j} z_{j-1} + \frac{h}{a_j} z_1 z_j \quad (1 < j < l) \\ g_l(\mathbf{z}) &= \frac{b}{a_l} z_l^2 + \frac{d+ls}{a_l} + \frac{le}{a_l} z_{l-1} + \frac{h}{a_l} z_1 z_l,\end{aligned}$$

where $a_k = b + d + ks + u + ke + h$ is the event rate of a cell with k ISs (see subsection 2.3.1).

From the probability generating function, we derive the infinitesimal generating function $\tilde{g}_k(\mathbf{z}) = a_k(g_k(\mathbf{z}) - z_k)$:

$$\tilde{g}_1(\mathbf{z}) = (b+h)z_1^2 - (b+h+d+s+u+e)z_1 + uz_2 + d + s + e$$
$$\tilde{g}_j(\mathbf{z}) = bz_j^2 - (b+h+d+js+u+je)z_j + uz_{j+1} + jez_{j-1} + hz_1z_j + d + js$$
$$\tilde{g}_l(\mathbf{z}) = bz_l^2 - (b+h+d+ls+le)z_l + lez_{l-1} + hz_1z_l + d + ls$$

and the infinitesimal generator $A = (a_{ij}) = a_i b_{ij}$, where $b_{ij} = \frac{\partial g_i(\mathbf{z})}{\partial z_j}\Big|_{\mathbf{z}=\mathbf{1}} - \delta_{ij}$:

$$A = \begin{pmatrix} b+h-d-s-u-e & u & & & & & & \\ h+2e & b-d-2s-u-2e & u & & & & & \\ h & 3e & b-d-3s-u-3e & u & & & & \\ \vdots & & & \ddots & & & & \\ h & & & je & b-d-js-u-je & u & & \\ \vdots & & & & & \ddots & & \\ h & & & & & le & b-d-ls-le & \end{pmatrix}$$

2.5.2 Results: Time to threshold

To extend the ordinary differential equations given in subsection 2.3.3 to $x = 0$, we observe that

$$\begin{aligned}\frac{d\varphi^m(x)}{dx} &= \frac{d}{dx}\mathbb{E}\left(e^{iW^m x}\right) = \frac{d}{dx}\int_0^\infty e^{itx} f^m(t)dt \overset{\text{Leibniz}}{=} \int_0^\infty \frac{\partial}{\partial x}\left[e^{itx} f^m(t)\right] dt \\ &= \int_0^\infty ite^{itx} f^m(t)dt = \mathbb{E}\left(iW^m e^{iW^m x}\right),\end{aligned}$$

where $f^m(t)$ is the probability distribution of W^m, and so $\frac{d\varphi^m(x)}{dx}\Big|_{x=0} = i\mathbb{E}\left(W^m\right) = iu_m$, where u_m is the m-th component of the scaled right eigenvector $\mathbf{u}$ to the eigenvalue λ_0 of the infinitesimal generator A.

Therefore, the ordinary differential equation system for $\varphi^m(x)$, $m \in \{1, \ldots, l\}$, is

$$\begin{aligned}
\frac{d\varphi^1(x)}{dx} &= \frac{1}{\lambda_0 x}\,[\,(b+h)(\varphi^1(x))^2 - (b+h+d+s+u+e)\varphi^1(x) \\
&\qquad + u\varphi^2(x) + d + s + e\,] \\
\frac{d\varphi^j(x)}{dx} &= \frac{1}{\lambda_0 x}\,[\,h\varphi^1(x)\varphi^j(x) + b(\varphi^j(x))^2 - (b+h+d+js+u+je)\varphi^j(x) \\
&\qquad + u\varphi^{j+1}(x) + je\varphi^{j-1}(x) + d + js\,] \quad (1 < j < l) \\
\frac{d\varphi^l(x)}{dx} &= \frac{1}{\lambda_0 x}\,[\,h\varphi^1(x)\varphi^l(x) + b(\varphi^l(x))^2 - (b+h+d+ls+le)\varphi^l(x) \\
&\qquad + le\varphi^{l-1}(x) + d + ls\,]
\end{aligned}$$

if $x \neq 0$, and

$$\left.\frac{d\varphi^m(x)}{dx}\right|_{x=0} = iu_m \quad \text{for } m \in \{1, \ldots, l\}$$

if $x = 0$, with

$$\varphi^m(0) = 1 \text{ for } m \in \{1, \ldots, l\}.$$

3. Estimating the Fitness Effect of an Insertion Sequence

Manuel Bichsel, Andrew D. Barbour, Andreas Wagner; *Journal of Mathematical Biology*, 2013, 66(1-2):95–114

Abstract

Since its discovery, mobile DNA has fascinated researchers. In particular, many researchers have debated why insertion sequences persist in prokaryote genomes and populations. While some authors think that insertion sequences persist only because of occasional beneficial effects they have on their hosts, others argue that horizontal gene transfer is strong enough to overcome their generally detrimental effects. In this study, we model the long-term fate of a prokaryote cell population, of which a small proportion of cells has been infected with one insertion sequence per cell. Based on our model and the distribution of IS*5*, an insertion sequence for which sufficient data is available in 525 fully sequenced proteobacterial genomes, we show that the fitness cost of insertion sequences is so small that they are effectively neutral or only slightly detrimental. We also show that an insertion sequence infection can persist and reach the empirically observed distribution if the rate of horizontal gene transfer is at least as large as the fitness cost, and that this rate is well within the rates of horizontal gene transfer observed in nature. In addition, we show that the time needed to reach the observed prevalence of IS*5* is unrealistically long for the fitness cost and horizontal gene transfer rate that we computed. Occasional beneficial effects may thus have played an important role in the fast spreading of insertion sequences like IS*5*.

3.1 Introduction

Bacterial insertion sequences (ISs) are the simplest form of autonomous mobile DNA. They are short (800–2500 bp) DNA sequences typically consisting of one open reading frame that codes for the enzyme transposase which is needed for transposition. The open reading frame is flanked by short terminal inverted repeats which serve as recognition sites for the transposase. This enzyme usually excises the IS and inserts it elsewhere in the genome (conservative transposition), but occasionally it replicates the IS during this transposition process (replicative transposition) [Chandler and Mahillon, 2002]. An IS may get lost from a genome through excision. ISs have been assigned numbers, roughly in the order of their discovery: e.g. IS*1*A, IS*5*, IS*630*. Based on their internal structure and the inverted repeats, all ISs have been classified into 20 different families [Chandler and Mahillon, 2002, Mahillon et al., 2009]. The focus of our study, IS*5*, belongs to a rather heterogeneous family of ISs that is widely distributed among bacteria and archaea.

IS*5* and all other ISs are inherited through vertical transmission. But they can also be horizontally transmitted by horizontal gene transfer (HGT) between prokaryotes, i.e. by natural transformation, by transduction through phages, and by conjugation through plasmids. The reported rates of transposition, excision and HGT are typically very low. Table 3.1 provides an overview over these rates.

Event		Rates	Sources
Transposition	Conservative	10^{-7} – 10^{-4}	[Kleckner, 1989], [Chandler and Mahillon, 2002]
Excision		10^{-10}	[Kleckner, 1989]
HGT	Transformation	10^{-6} – 10^{-3}	[Williams et al., 1996]
	Transduction	10^{-8}	[Jiang and Paul, 1998]
	Conjugation	10^{-6} – 10^{-5}	[Dahlberg et al., 1998]

Table 3.1: Transposition, excision, and HGT rates reported by different authors. Rates have been converted into events per cell or IS and generation.

Due to their transposition activity and the deletions, insertions and inversions through homologous recombination that are possible if more than one IS is present in a genome [Galas and Chandler, 1989, Kleckner, 1989, Schneider and Lenski, 2004], ISs pose a potential

threat to their hosts, although occasional beneficial effects have also been reported [Hall, 1999, Schneider and Lenski, 2004]. Besides acting on their own, two ISs can also form a composite transposon, which consists of two copies of an IS that flank intermediary genes and transpose synchronously, thereby mobilising the intermediary genes. In this way, ISs are involved in transferring genes that confer resistance to antibiotics [Berg, 1989, Kleckner, 1989], genes that encode toxins [So and McCarthy, 1980], or genes with new metabolic functions [Top and Springael, 2003]. On the one hand, ISs therefore help to spread antibiotic resistance among pathogens and pose a public health threat, but on the other hand, ISs are also valuable tools used in genetic engineering.

While most authors agree that harboring ISs in the genome is in general detrimental to the cell, there is disagreement about whether ISs persist because they are occasionally beneficial to their hosts [Blot, 1994, Shapiro, 1999, Schneider and Lenski, 2004] or because HGT is sufficiently strong to overcome selection against ISs due to their detrimental effects on their hosts [Dawkins, 1976, Doolittle and Sapienza, 1980, Orgel and Crick, 1980, Charlesworth et al., 1994, Nuzhdin, 1999].

In an earlier study, we used a stochastic, branching process model to show that even purely detrimental ISs can invade a host cell population and persist, provided that the HGT rate is larger than the fitness cost caused by the IS [Bichsel et al., 2010]. Based on our model, we showed that most IS infections do not persist and die out very quickly. Those infections that do persist, take a very long time to reach noticeable population sizes. While the branching process model is well suited to model the initial phase of an IS infection, it does not allow for interactions between cells, and is not useful for modeling the long-term effects of an IS infection. In this study, we therefore use a deterministic model based on a system of ordinary differential equations to examine whether purely detrimental ISs can persist. We then determine how large a fitness cost of an IS and a HGT rate would be needed to obtain the IS count distribution we observe in bacterial genomes. In doing so, we focus on the largely proteobacterial insertion sequence IS*5*, the only IS for which sufficient data is available. Because very similar IS count distributions have been observed in many other ISs [Sawyer et al., 1987, Wagner, 2006, Touchon and Rocha, 2007], we presume that our results are qualitatively comparable to those that would be obtained for other ISs.

3.2 Data, Model, and Methods

3.2.1 Data

We obtained the genome sequences of 1447 fully sequenced prokaryote genomes from 542 genera that were available at NCBI on September 1, 2011 [NCBI, 2010]. We also obtained the sequences of one representative IS from each of the 20 known IS families from the IS Finder database [Mahillon et al., 2009]. We then used IScan [Wagner et al., 2007] to search the 1447 genome sequences (only chromosomes, no plasmids) for these 20 representative IS sequences. For later analysis, we needed independent IS count observations. We were therefore interested in ISs that occur in many genera. Of all 20 ISs, only 3 occur in more than 10 different genera. And of these 3 ISs, only IS*5* occurs often enough in these genera so that a random sample of one genome per genus contains on average more than 10 infected genomes. To get more dependable results in our statistical analysis, we therefore focused on IS*5*. It turned out that IS*5* (as most of the other 20 ISs we examined) can be found mainly in genomes from proteobacteria: only 4 of 58 infected genomes do not belong to proteobacteria. We thus restricted our IS*5* count analysis to proteobacteria. In our data set, this phylum consists of 525 genomes in 180 genera, where we have added Shigella to the genus Escherichia because of their well-known close phylogenetic relationship [Lan and Reeves, 2002].

3.2.2 Model Design

Figure 3.1 shows the design of our model, in which we assume an uninfected prokaryote host cell population living at carrying capacity K, where K is a cell population density. The prokaryote cells live in a well-mixed bulk environment, and their normalized population density is given by $Z_0 = D_0/K$, where D_0 is their density. The change of Z_0 over time is governed by the logistic equation $\dot{Z}_0 = r(1 - Z_0)Z_0$, with the base population growth rate r. We set $r = 1$, so that one time unit corresponds to the doubling time during the early exponential growth phase, and we take this as the generation time of a cell. At the begin of an IS infection, each cell of a very small proportion of cells (e.g. 10^{-6}) is infected with one IS in its genome. We then model the spread of the IS infection through the host cells

and compute the equilibrium distribution of the IS count per prokaryotic cell genome. To do so, we use a system of ordinary differential equations for the normalized cell densities $Z_k = D_k/K$, where D_k is the density of cells carrying k ISs in their genome. To keep the computation numerically tractable, we limit the maximal number of ISs per infected cell to $l = 60$. This is not a strong limitation, because only few genomes harbor more than 60 ISs, as other authors have reported [Sawyer et al., 1987, Wagner, 2006, Touchon and Rocha, 2007]. Furthermore, we show in the Results section that no genome in our data set contains more than 60 copies of IS*5*, the focus of our interest. Besides the base population growth rate r, our model contains the following rate parameters: the base fitness effect s of one IS, the base replicative transposition rate u of one IS, the excision rate e per IS, and the HGT rate h.

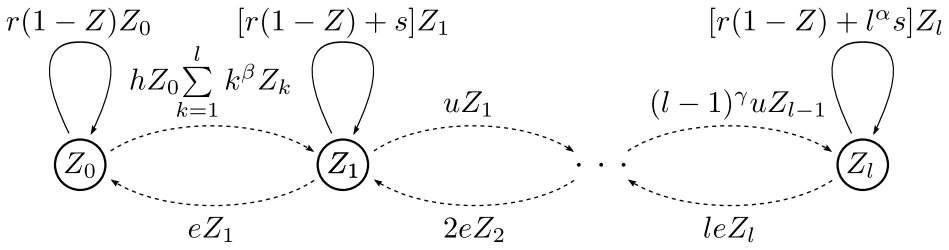

Figure 3.1: Model design. $Z_k = D_k/K$ is the normalized density of cells with $k \in \{0, \ldots, l\}$ ISs, where D_k is the density of cells with k ISs, and K is the carrying capacity; $r = 1$ is the base growth rate per uninfected cell; s is the base fitness effect of one IS; u is the base replicative transposition rate of one IS; e is the excision rate per IS; h is the HGT rate; $l = 60$ is the maximal IS count per genome; $\alpha, \beta, \gamma \in \{0, 1, 2\}$ are power function exponents that control the increase of the fitness effect, of the HGT rate, and of the replicative transposition rate with increasing IS count per cell. All rates ar per time unit. Because $r = 1$, one time unit corresponds to the doubling time during the exponential cell growth phase. Solid lines indicate a change in the total cell density, and dashed lines indicate a change only in the normalized density distribution of the cells with different IS counts in their genome.

We allow for a nonlinear impact of an increasing IS count per cell on the cell's fitness, its infectiousness to other cells, and its total replicative transposition rate. To this end, we model the fitness effect, the HGT rate and the total replicative transposition rate of all ISs in a cell as a power function of the cell's IS count, with exponents α, β and γ, respectively. We choose $\alpha, \beta, \gamma \in \{0, 1, 2\}$, where an exponent of 0 reflects independence of the rate from the cell's total IS count, an exponent of 1 reflects linear dependence, and an exponent of 2 reflects quadratic dependence of the rate from the cell's total IS count. This is equivalent to a diminishing (exponent 0), a constant (exponent 1) and an increasing (exponent 2) effect

per IS of an increasing IS count on the rate. For simplicity, we let the total excision rate increase linearly with the cell's IS count, i.e. we assume that ISs are excised independently of each other.

Our data suggest that the number of infected cells in a population stays low compared to the total number of cells (see figure 3.2). This has also been observed before [Wagner, 2006, Touchon and Rocha, 2007]. To simplify our model, we therefore assume that infected cells are surrounded by uninfected cells only, and that no HGT occurs between infected cells. Furthermore, we assume that during HGT only one IS gets copied from an infected to an uninfected cell. This is justified by the observation that during transformation and transduction typically only small DNA fragments are transferred from one cell to another, and that ISs on a plasmid transferred during conjugation must first be inserted into the chromosome [Madigan et al., 2009, p. 297ff].

3.2.3 Model Analysis

Based on our model design shown in figure 3.1, we describe the dynamics of an IS infection with the following system of ordinary differential equations, where $Z = \sum_{k=0}^{l} Z_k \geq 0$:

$$\begin{aligned}
\dot{Z}_0 &= r(1-Z)Z_0 - h\,Z_0 \sum_{k=1}^{l} k^\beta Z_k + eZ_1 \\
\dot{Z}_1 &= [r(1-Z)+s]\,Z_1 + h\,Z_0 \sum_{k=1}^{l} k^\beta Z_k - uZ_1 - eZ_1 + 2eZ_2 \\
\dot{Z}_2 &= [r(1-Z)+2^\alpha s]\,Z_2 + uZ_1 - 2^\gamma uZ_2 - 2eZ_2 + 3eZ_3 \\
&\vdots \\
\dot{Z}_j &= [r(1-Z)+j^\alpha s]\,Z_j + (j-1)^\gamma uZ_{j-1} - j^\gamma uZ_j - jeZ_j + (j+1)eZ_{j+1} \\
&\vdots \\
\dot{Z}_l &= [r(1-Z)+l^\alpha s]\,Z_l + (l-1)^\gamma uZ_{l-1} - leZ_l
\end{aligned} \tag{3.1}$$

This system has two obvious equilibrium solutions: the first one is $Z_0 = Z_1 = \ldots = Z_l = 0$, i.e. population extinction, and the second one is $Z_0 = 1$ and $Z_1 = \ldots = Z_l = 0$, i.e. IS extinction. We are more interested in equilibria where not all Z_k for $k \in \{1, \ldots, l\}$ vanish. In that case $Z > 0$, and using the proportions $p_k = Z_k/Z$ and their derivatives with

respect to time,

$$\dot{p_k} = \frac{\dot{Z_k}}{Z} - \frac{Z_k \cdot \dot{Z}}{Z^2} = \frac{1}{Z}\dot{Z_k} - p_k \frac{1}{Z} \sum_{j=0}^{l} \dot{Z_j} = \frac{1}{Z}\dot{Z_k} - p_k \left(r(1-Z) + s \sum_{j=1}^{l} j^\alpha p_j \right),$$

we define a new system of ordinary differential equations for p_k ($k \in \{0, \ldots, l\}$) and for Z:

$$\begin{aligned}
\dot{p_0} &= -h\, p_0\, Z \sum_{k=1}^{l} k^\beta p_k + e p_1 - s\, p_0 \sum_{k=1}^{l} k^\alpha p_k \\
\dot{p_1} &= s p_1 + h\, p_0\, Z \sum_{k=1}^{l} k^\beta p_k - u p_1 - e p_1 + 2 e p_2 - s\, p_1 \sum_{k=1}^{l} k^\alpha p_k \\
\dot{p_2} &= 2^\alpha s p_2 + u p_1 - 2^\gamma u p_2 - 2 e p_2 + 3 e p_3 - s\, p_2 \sum_{k=1}^{l} k^\alpha p_k \\
&\vdots \\
\dot{p_j} &= j^\alpha s p_j + (j-1)^\gamma u p_{j-1} - j^\gamma u p_j - j e p_j + (j+1) e p_{j+1} - s\, p_j \sum_{k=1}^{l} k^\alpha p_k \\
&\vdots \\
\dot{p_l} &= l^\alpha s p_l + (l-1)^\gamma u p_{l-1} - l e p_l - s\, p_l \sum_{k=1}^{l} k^\alpha p_k
\end{aligned} \tag{3.2}$$

and

$$\dot{Z} = \sum_{k=0}^{l} \dot{Z_k} = r(1-Z)Z + s\, Z \sum_{k=1}^{l} k^\alpha p_k = \left(r(1-Z) + s \sum_{k=1}^{l} k^\alpha p_k \right) Z. \tag{3.3}$$

Besides setting $r = 1$, we set the replicative transposition rate u and the excision rate e to one of two fixed parameter sets that together cover a range of realistic rates (see table 3.1). In the main text, we use $(u, e) = (10^{-7}, 10^{-10})$, and in the appendix we use $(u, e) = (10^{-9}, 10^{-11})$. To solve the system (3.2, 3.3), we define $\mathbf{p} = (p_0, \ldots, p_l)^T$, $S_\alpha(\mathbf{p}) = \sum_{k=1}^{l} k^\alpha p_k$, $S_\beta(\mathbf{p}) = \sum_{k=1}^{l} k^\beta p_k$, and the HGT parameter $H(\mathbf{p}, Z) = h\, S_\beta(\mathbf{p})\, Z$. Observe that $H \geq 0$. The differential equations for $p_0, \ldots, p_l$ in (3.2) can now be written in vector notation as

$$\dot{\mathbf{p}} = \mathbf{M}(s, H(\mathbf{p}, Z)) \cdot \mathbf{p} - s\, S_\alpha(\mathbf{p})\, \mathbf{p} \tag{3.4}$$

where

$$\mathbf{M}(s,H) = \begin{pmatrix} -H & e & & & & & & \\ H & s-u-e & 2e & & & & & \\ & u & 2^\alpha s - 2^\gamma u - 2e & 3e & & & & \\ & & \dots & \dots & \dots & & & \\ & & & (j-1)^\gamma u & j^\alpha s - j^\gamma u - je & (j+1)e & & \\ & & & & \dots & \dots & \dots & \\ & & & & & (l-1)^\gamma u & l^\alpha s - le \end{pmatrix}$$

We get again the IS extinction equilibrium for $\mathbf{p} = \mathbf{e}_0 = (1, 0, \dots, 0)^T$, because $S_\alpha(\mathbf{e}_0) = S_\beta(\mathbf{e}_0) = H(\mathbf{e}_0, Z) = 0$ for any $Z > 0$, and therefore $\mathbf{M}(s,0) \cdot \mathbf{e}_0 = \mathbf{0}$, so that $\mathbf{M}(s,0) \cdot \mathbf{p} - s\, S_\alpha(\mathbf{p})\, \mathbf{p} = \mathbf{0}$. For all other equilibrium solutions $(\mathbf{p}, Z)$ of (3.2, 3.3) that may exist, $H = H(\mathbf{p}, Z)$ and $\lambda = s\, S_\alpha(\mathbf{p})$ must fulfill $\mathbf{M}(s,H) \cdot \mathbf{p} = \lambda\, \mathbf{p}$. We are therefore looking for non-negative eigenvectors of the matrix $\mathbf{M}(s,H)$ for $H > 0$ ($H = 0$ is not interesting, because it implies $Z_1 = \ldots = Z_l = 0$).

$\mathbf{M}(s,H)$ for $H > 0$ is a Metzler-Leontief matrix, i.e. $(\mathbf{M})_{ij} \geq 0$ for $i \neq j$ [Seneta, 1981, p. 45]. In addition, $\mathbf{M}$ is irreducible. Therefore, for any choice of $H > 0$, there exists an eigenvalue $\tau \in \mathbb{R}$ such that $\tau > \mathrm{Re}(\mu)$ for all other eigenvalues μ of $\mathbf{M}$, and there exists a unique (up to multiples), strictly positive eigenvector $\mathbf{q}$ associated with τ. $\mathbf{q}$ can be normed so that $\|\mathbf{q}\|_1 = 1$. Furthermore,

1. if $\mathbf{M}(s,H) \cdot \mathbf{p} = \eta\, \mathbf{p}$ for a specific eigenvector $\mathbf{p}$ with $\sum_{k=0}^{l} p_k = 1$, then $(1, \dots, 1) \cdot \mathbf{M}(s,H) \cdot \mathbf{p} = \eta\, (1, \dots, 1) \cdot \mathbf{p} = \eta$,
2. $(1, \dots, 1) \cdot \mathbf{M}(s,H) \cdot \mathbf{p} = s\, S_\alpha(\mathbf{p}) = \lambda$ for all proportion vectors $\mathbf{p}$ (see the differential equations (3.2) for $\mathbf{p}$),

and therefore $\tau = \lambda = s\, S_\alpha(\mathbf{q})$.

We now have

$$\dot{\mathbf{q}} = \mathbf{M}(s, H(\mathbf{q}, Z)) \cdot \mathbf{q} - s\, S_\alpha(\mathbf{q})\, \mathbf{q} = 0,$$

and therefore, if we set $Z = 1 + \frac{s}{r} S_\alpha(\mathbf{q})$, so that $\dot{Z}(\mathbf{q}) = [r(1 - Z) + s\, S_\alpha(\mathbf{q})]\, Z = 0$, the pair $(\mathbf{q}, Z)$ is an equilibrium solution of the system (3.2, 3.3) for the proportions $\mathbf{p}$ and the total population size Z.

Note that it is hard to compute an equilibrium solution based directly on h, β, and s, because one then has to solve the differential equation system (3.2, 3.3). But it is much easier to algebraically compute an equilibrium solution of (3.2, 3.3) for a given pair (s, H) with $H > 0$ and then to find values of $h = h_\beta$ for any $\beta \in \{0, 1, 2\}$. These are the required computational steps:

1. Compute the unique eigenvector $\mathbf{q}$ with $\|\mathbf{q}\|_1 = 1$ that corresponds to the (real) eigenvalue τ with the largest real part of the matrix $\mathbf{M}(s, H)$.
2. Set $Z = 1 + \frac{s}{r} S_\alpha(\mathbf{q}) = 1 + \frac{s}{r} \sum_{k=1}^{l} k^\alpha q_k$.
3. Compute $h_\beta = \frac{H}{S_\beta(\mathbf{q})\, Z} = \frac{H}{\sum_{k=1}^{l} k^\beta q_k\, Z}$ for $\beta \in \{0, 1, 2\}$.

To assess the local stability of the equilibrium distribution $(q_0, \ldots, q_l)$, we first calculate $Z_j = q_j\, Z$ for $j \in \{0, \ldots, l\}$. We then compute the eigenvalues of the Jacobian matrix $J = \left(\frac{\partial f_i(Z_0, \ldots, Z_l)}{\partial Z_j}\right)_{i,j \in \{0,\ldots,l\}}$ at the specific values of (s, h) and (α, β, γ) used to compute the equilibrium. Here, $f_i(Z_1, \ldots, Z_l)$ is the right-hand side of the differential equation for Z_i in the system (3.1). The equilibrium is locally stable if the real parts of all eigenvalues are negative.

The global stability of the equilibrium is much more difficult to establish. We confine ourselves to check whether the equilibrium is reached, starting from an initial cell population at carrying capacity infected with a small proportion of cells harboring one IS in their genome. To do so, we numerically solve the system (3.1) with the values of (s, h) and (α, β, γ) used to compute the equilibrium. We have chosen the values 10^{-9}, 10^{-6}, and 10^{-3} as proportions of initially infected cells.

To find values for (α, β, γ) and (s, h) that lead to the best approximation of an observed IS*5* count distribution by our theoretical IS count distribution, we use a maximum likelihood method and compute maximum likelihood estimates of (α, β, γ) and (s, h). We start by defining the likelihood function L. Given the observed IS counts $(c_0, \ldots, c_l)$ and the predicted IS count distribution $(q_0, \ldots, q_l)$ based on the parameters α, γ, s and H, the

likelihood function is given by

$$L(\alpha, \gamma, s, H) = q_0^{c_0} \cdot \ldots \cdot q_l^{c_l},$$

and its (natural) logarithm is

$$\ln\left(L(\alpha, \gamma, s, H)\right) = c_0 \cdot \ln(q_0) + \ldots + c_l \cdot \ln(q_l).$$

Because we can only numerically compute the vector of proportions $\mathbf{q}$ based on the parameters α, γ, s, and H, and because we cannot derive $\mathbf{q}$ in analytical form, we use the Nelder-Mead method [Nelder and Mead, 1965] to find the maximum log-likelihood in the parameter space (s, H) for all pairs $(\alpha, \gamma) \in \{0, 1, 2\}^2$. Having found the maximum likelihood estimates $\hat{s}$ and $\hat{H}$ for the combination $(\hat{\alpha}, \hat{\gamma})$ of α and γ that maximises the likelihood function L, we then obtain the maximum likelihood estimate $\hat{h} = \hat{h_\beta}$ for all values of $\beta \in \{0, 1, 2\}$ by following the three computational steps described above, replacing s by $\hat{s}$ and H by $\hat{H}$ throughout.

For four specific maximum likelihood estimates $(\hat{s}, \hat{H})$, based on four different exponent pairs (α, γ) we then use the bootstrap method with 1000 artificially generated data sets to show the association between the fitness effect s and the HGT rate h, and to compute the corresponding 95%-confidence intervals [Efron and Tibshirani, 1994, p. 170].

For the numerical analysis, we use Mathematica 8.0.0 [Wolfram, 2003].

3.3 Results

3.3.1 The IS*5* count distribution in proteobacterial cells is L-shaped

Figure 3.2 shows the IS*5* count distribution based on 525 fully sequenced, proteobacterial genomes. We have generated 1000 random samples of genomes. Each sample consists of 180 genomes, one randomly chosen genome per proteobacterial genus. We then counted how many genomes per random sample contained 0 ISs, 1-5 ISs, ..., 16-20 ISs, or more than 20 ISs. Figure 3.2 shows the mean number of genomes per IS count bin over all 1000 samples, together with the 10th and 90th percentile. Averaging over 1000 random samples

provides us with an approximation of the real IS*5* count distribution over proteobacterial genera. Furthermore the 1000 random samples provide insight into the uncertainty about the real IS*5* count distribution, and about the resulting uncertainty in determining model parameters.

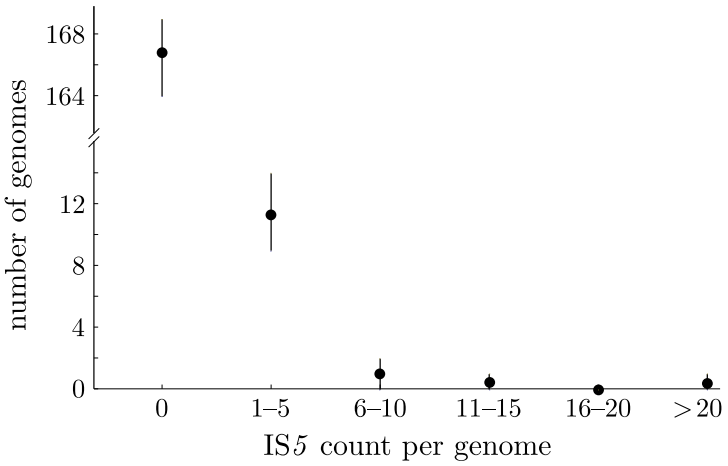

Figure 3.2: IS*5* count distribution of 1000 random samples from 525 proteobacterial genomes, each sample containing 180 genomes. Different IS counts have been collected into bins. Dots mark the mean number of genomes in the corresponding bin, and the lower and upper ends of the vertical lines mark the 10th and 90th percentile of the number of genomes, respectively. Note the discontinuous scale on the vertical axis.

As can be seen in the figure, the IS*5* count distribution in proteobacterial cells is strongly L-shaped. An overwhelming majority of genomes, namely 92.7%, does not contain any IS*5* copies, a small fraction of genomes contains up to 10 or 15 copies, and only few genomes contain more than 15 copies, although there are proteobacterial genomes with higher IS*5* counts. The *Pseudomonas syringae* tomato DC3000 genome and all three *Xanthomonas oryzae* genomes in our data set contain more than 40 copies of IS*5*, where *Xanthomonas oryzae* MAFF 311018 has the highest count of 54 IS*5* copies.

3.3.2 The HGT rate has to be larger than the fitness cost of IS*5* for an IS infection to reach the observed IS*5* count distribution in equilibrium

We set the replicative transposition rate u and the excision rate e to $(u, e) = (10^{-7}, 10^{-10})$, which is in the range of values provided in table 3.1. Note that table 3.1 reports only the conservative transposition rate, and we assume that the replicative transposition rate is

a few orders of magnitude smaller [Tavakoli and Derbyshire, 2001]. In the appendix, we present analogous results using a different set $(u, e) = (10^{-9}, 10^{-11})$ of parameters.

Next, we compute the maximum likelihood estimates of the fitness effect s and the HGT parameter H for all 9 possible combinations of the fitness effect exponent α and the replicative transposition exponent γ. We do so for all 1000 random samples of size 180 from 525 proteobacterial genomes. To identify in each sample those models that do not fit the data significantly worse than the best model for the sample, we follow an argument of Sawyer et al. [Sawyer et al., 1987] and take in each sample the model with the highest log-likelihood as a proxy for the model with free (and continuous) parameters α and γ. As a consequence, all our models with specific, fixed α and γ become nested within this proxy model that is considered to have two additional degrees of freedom. We can then apply the likelihood-ratio test in each sample to compare the proxy model with all other models, using a χ^2 distribution with two degrees of freedom. Therefore, on a 5% significance level, models whose log-likelihood is not by more than $\chi^2_{0.05,2}/2 = 3.0$ units lower than the log-likelihood of the best model of the sample, fit observed data not significantly worse than this best model. Applied to our data, we find that the exponent combinations $(\alpha, \gamma) = (0, 1)$ and $(\alpha, \gamma) = (1, 1)$ lead to the best fit in 359 and 346 of the 1000 samples, respectively. Based on our criterium for the log-likelihood described above, we find that the exponent combinations $(\alpha, \gamma) = (0, 1)$ (in all 1000 samples), $(\alpha, \gamma) = (2, 2)$ (in 990 samples), $(\alpha, \gamma) = (1, 2)$ (in 957 samples), and $(\alpha, \gamma) = (1, 1)$ (in 951 samples) lead in over 90% of all samples to fits that are not significantly worse than the best fit in each sample. These findings for γ suggest that if the assumptions of the model are correct, the transposition rate per IS*5* copy does not decrease with increasing IS*5* count per genome. The fitness exponent parameter α does not show a clear distribution pattern, i.e. all its possible values (0, 1, and 2) can lead in over 90% of all samples to a fit that is not significantly worse than the best fit in each sample. Our data does therefore not allow to draw conclusions about possible interactions between IS*5* copies in influencing the fitness of a host cell.

Based on the maximum likelihood estimates of the fitness effect, $\hat{s}$, and of the HGT parameter, $\hat{H}$, we can compute the total population size $Z = 1 + \frac{\hat{s}}{r} S_\alpha(\mathbf{q})$ in equilibrium, as well as the maximum likelihood estimate of the HGT rate, $\hat{h} = \frac{\hat{H}}{S_\beta(\mathbf{q})\, Z}$, which depends

on our choice of the HGT exponent β. Table 3.2 shows for the four exponent pairs of $(\alpha, \gamma) \in \{(0,1),(1,1),(1,2),(2,2)\}$ that we found above the quartiles of the maximum likelihood estimates of s and of h for different choices of $\beta \in \{0,1,2\}$. We show only those HGT rates which lead to a stable equilibrium that can be reached by starting with a small proportion of infected cells (between 10^{-9} and 10^{-3}) carrying one copy of IS*5*.

(α,γ)	Quart.	$\hat{s}$	$\hat{h_0}$	$\hat{h_1}$	$\hat{h_2}$
$(0,1)$	Q1	$-1.6 \cdot 10^{-7}$	$1.1 \cdot 10^{-7}$	–	–
	Q2	$-1.3 \cdot 10^{-7}$	$1.3 \cdot 10^{-7}$	–	–
	Q3	$-1.1 \cdot 10^{-7}$	$1.6 \cdot 10^{-7}$	–	–
$(1,1)$	Q1	$-2.8 \cdot 10^{-8}$	$3.8 \cdot 10^{-8}$	$7.3 \cdot 10^{-9}$	–
	Q2	$-1.8 \cdot 10^{-8}$	$5.5 \cdot 10^{-8}$	$1.8 \cdot 10^{-8}$	–
	Q3	$-7.3 \cdot 10^{-9}$	$6.7 \cdot 10^{-8}$	$2.8 \cdot 10^{-8}$	–
$(1,2)$	Q1	$-2.9 \cdot 10^{-8}$	$7.7 \cdot 10^{-8}$	$1.9 \cdot 10^{-8}$	–
	Q2	$-2.3 \cdot 10^{-8}$	$8.2 \cdot 10^{-8}$	$2.3 \cdot 10^{-8}$	–
	Q3	$-1.9 \cdot 10^{-8}$	$8.8 \cdot 10^{-8}$	$2.9 \cdot 10^{-8}$	–
$(2,2)$	Q1	$-3.1 \cdot 10^{-9}$	$6.3 \cdot 10^{-8}$	$1.8 \cdot 10^{-8}$	$7.9 \cdot 10^{-10}$
	Q2	$-1.6 \cdot 10^{-9}$	$6.5 \cdot 10^{-8}$	$2.3 \cdot 10^{-8}$	$1.6 \cdot 10^{-9}$
	Q3	$-7.9 \cdot 10^{-10}$	$6.8 \cdot 10^{-8}$	$2.9 \cdot 10^{-8}$	$3.1 \cdot 10^{-9}$

Table 3.2: For a replicative transposition rate $u = 10^{-7}$ and an excision rate $e = 10^{-10}$, the table shows the four most frequent exponent pairs $(\alpha, \gamma) \in \{0,1,2\}^2$ that lead to model fits of the IS*5* count distribution that are not significantly worse than the best fit. For each pair (α, γ), the quartiles (Q1, Q2, Q3, where Q2 is the median) of the maximum likelihood estimates of the fitness effect s and of the HGT rate h_β for different scaling exponents $\beta \in \{0,1,2\}$ of the HGT rate are reported. Only HGT rates that lead to stable equilibria are shown. Observe that Q1 in $\hat{s}$ corresponds to Q3 in $\hat{h_\beta}$ and vice versa.

Table 3.2 shows that $\hat{s} < 0$ for IS*5*, i.e. that IS*5* is generally detrimental. The table also shows that $\beta \leq \alpha$ is needed to reach the equilibrium, starting with a small proportion of infected cells. In that case, $\hat{h_\beta} \geq |\hat{s}|$. If, on the other hand, $\beta > \alpha$, then $\hat{h_\beta} < |\hat{s}|$ (not shown). The IS*5* infection, starting with cells carrying one copy of IS*5* only, will then die out, because HGT is not strong enough to overcome the negative fitness effect caused by even only one IS*5* copy per genome. Therefore, for an IS infection to spread, persist, and reach the observed IS*5* count distribution, the increase in the infectiousness of a cell with increasing IS count must be smaller than the simultaneous increase in the total fitness cost.

Figure 3.3 shows for each of the four exponent pairs $(\alpha, \gamma) \in \{(0,1),(1,1),(1,2),(2,2)\}$ an example of the predicted equilibrium distribution based on the maximum likelihood

estimates $\hat{s}$ and $\hat{H}$, together with an observed IS*5* count distribution based on a sample that led to the best model fit with the chosen pair (α, γ) and the maximum likelihood estimates $\hat{s}$ and $\hat{H}$. The four pairs (α, γ) cover all possible fitness effect exponents $\alpha \in \{0, 1, 2\}$, and they lead to a wide range of the estimated fitness effect $\hat{s}$ (compare with table 3.2). We truncate the computed distribution at $l = 60$ IS copies per genome. The bin with 60 copies per genome therefore represents all genomes with *at least* 60 copies in the computed distribution. (The highest IS*5* count in all proteobacterial genomes is 54 and therefore well below $l = 60$.)

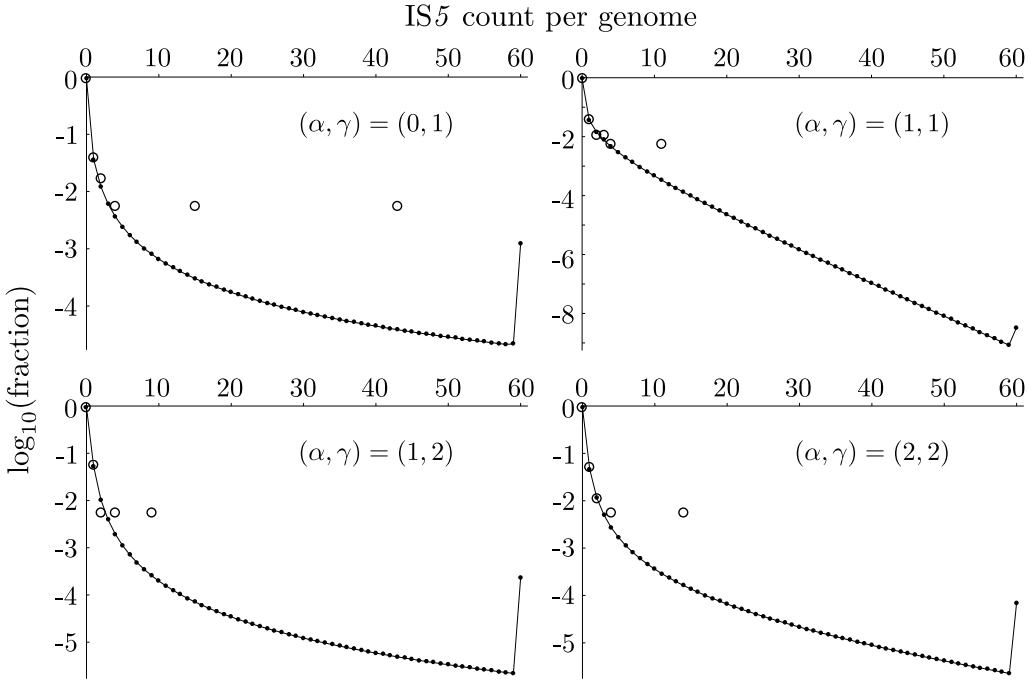

Figure 3.3: Observed (large circles) and predicted (small dots, connected by a solid line) IS*5* count distributions for $(\alpha, \gamma) \in \{(0, 1), (1, 1), (1, 2), (2, 2)\}$. In the predicted IS count distribution, the IS count per genome has been limited to $l = 60$ copies of IS*5*. Note the logarithmic scale on the vertical axis. $\log_{10}(1/180) \approx -2.3$, i.e. each large circle at $\log_{10}(\text{fraction}) = -2.3$ represents only one genome.

A conspicuous feature of the predicted IS count distributions is the sharp upward spike at the highest IS count. It stems from the truncation we imposed at $l = 60$ IS copies per genome. In a model with no upper bound for the IS count per genome, the distribution would drop monotonously. We have confirmed this by using higher IS count limits l and by observing that the spike in the highest IS count then gets smaller when we again apply the maximum likelihood method (results not shown).

To get an estimate of the time needed to approximately reach the equilibrium distribution, we compute the population dynamics of an IS*5* infection over time, again for the four exponent pairs $(\alpha, \beta) \in \{(0,1),(1,1),(1,2),(2,2)\}$ already used above, and with the corresponding maximum likelihood estimates of s and $h_\beta = h_0$. We choose to focus on $\beta = 0$, because HGT is tightly regulated and depends on several internal and external factors [Dröge et al., 1999], so that the infectiousness of a cell probably depends only very weakly or not at all on the cell genome's IS count. Our choice of the base population growth rate $r = 1$ means that one time unit corresponds to the doubling time during the early exponential growth phase of a cell population. We identify this doubling time with one cell generation and set one cell generation to one day [Gibbons and Kapsimalis, 1967, Savageau, 1983] for the purpose of this analysis. Our computations then show, on the one hand, that the time to reach 90% of the final prevalence of infected cells is very long if we start with an initial prevalence of 10^{-6} infected cells. It lies between $1.8 \cdot 10^7$ years for $(\alpha, \gamma) = (1,2)$ and $(\hat{s}, \hat{h_0}) = (-5.6 \cdot 10^{-8}, 1.1 \cdot 10^{-7})$ and $1.8 \cdot 10^{10}$ years for $(\alpha, \gamma) = (0,1)$ and $(\hat{s}, \hat{h_0}) = (-1.1 \cdot 10^{-7}, 1.1 \cdot 10^{-7})$. On the other hand, the predicted time needed for the population of infected cells only to approximately reach its final IS count distribution is much shorter. It lies between about 7'100 years for $(\alpha, \gamma) = (1,2)$ and 33'500 years for $(\alpha, \gamma) = (0,1)$. In the latter computation, we numerically solve the equation $\frac{1}{2}\sum_{j=1}^{l} |Z_j(t)/Z_{\text{inf}}(t) - Z_j^\star/Z_{\text{inf}}^\star| = 0.1$ for the time t, where $Z_{\text{inf}} = \sum_{j=1}^{l} Z_j$ and an asterisk (*) indicates the final normalized population densities. In the appendix, we show for $(\alpha, \gamma) = (0,1)$, $s = -1.1 \cdot 10^{-7}$ and $h_0 = 1.1 \cdot 10^{-7}$ the computed dynamics over time of a population of host cells infected with a fraction of 10^{-6} cells harboring one copy of an IS in their genome. We also demonstrate the effect of changing the transposition rate u, the IS excision rate e, the fitness effect s, and the HGT rate h_0 on the population dynamics and on the final IS count distribution.

We have computed the dynamics of the total host population size over time during an infection for each of the four exponent pairs (α, γ), using the same maximum likelihood estimates for s and h_0 as in the preceding paragraph. In all cases, the relative reduction in the normalized population density caused by the infection is negligible, between $4.7 \cdot 10^{-9}$ for $(\alpha, \gamma) = (1,1)$ and $8.2 \cdot 10^{-9}$ for $(\alpha, \gamma) = (1,2)$. This is expected, because the fitness

cost of IS*5* is generally small, as our computations show.

3.3.3 The maximum likelihood estimates of the HGT rate and of the fitness effect are highly correlated

Using 1000 random samples of 180 out of 525 proteobacterial genomes, one per genus, provides insights into the uncertainty about the real IS*5* count distribution, and the resulting uncertainty in determining exponent pairs (α, γ), fitness effect s, and HGT rate h. To also get information about the variation in the maximum likelihood estimates of s and h to be expected if our model were correct, and to gain some insight into the relationship between s and h, we use a bootstrap [Efron and Tibshirani, 1994, p. 170]. For each exponent pair $(\alpha, \beta) \in \{(0,1),(1,1),(1,2),(2,2)\}$, we choose the set of maximum likelihood estimates of s and H that led to the computed IS count distributions shown in figure 3.3. Based on each of the four computed IS count distributions, we then generated 1000 artificial data sets and determined for each data set the maximum likelihood estimates of s and h_0. Figure 3.4 shows the values of $\hat{h_0}$ versus $\hat{s}$. We again show only the graphs for $\beta = 0$, as in the preceding subsection, and we marked the pairs $(\hat{s}, \hat{h_0})$ in the 95% confidence interval of s with black dots, while pairs outside the confidence interval are marked with gray dots.

The 1000 bootstrapped pairs of $(\hat{s}, \hat{h_0})$ in figure 3.4 show an almost perfectly functional dependence between $\hat{h_0}$ and $\hat{s}$. To concisely describe this functional dependence, we have plotted the graph of the best fit of the shifted power function $\hat{h_0} = a(-\hat{s})^b + c$. As can be seen, the fit is very good, at least inside the 95% confidence interval. The functional dependence between $\hat{h_0}$ and $\hat{s}$ is almost linear if $\beta = \alpha$. We can understand this linear dependence by observing that from the first equation in the differential equation system (3.2) we get in equilibrium

$$h = \frac{ep_1 - sp_0 S_\alpha(\mathbf{p})}{p_0 Z S_\beta(\mathbf{p})} \approx -\frac{S_\alpha(\mathbf{p})}{S_\beta(\mathbf{p})} s. \tag{3.5}$$

Our model therefore suggests that the maximum likelihood estimate of the HGT rate depends very sensitively and almost exclusively on the maximum likelihood estimate of the fitness effect (and vice versa). This highlights the crucial role the HGT rate plays in surmounting the fitness cost of an IS and in allowing an IS to persist in a host cell population.

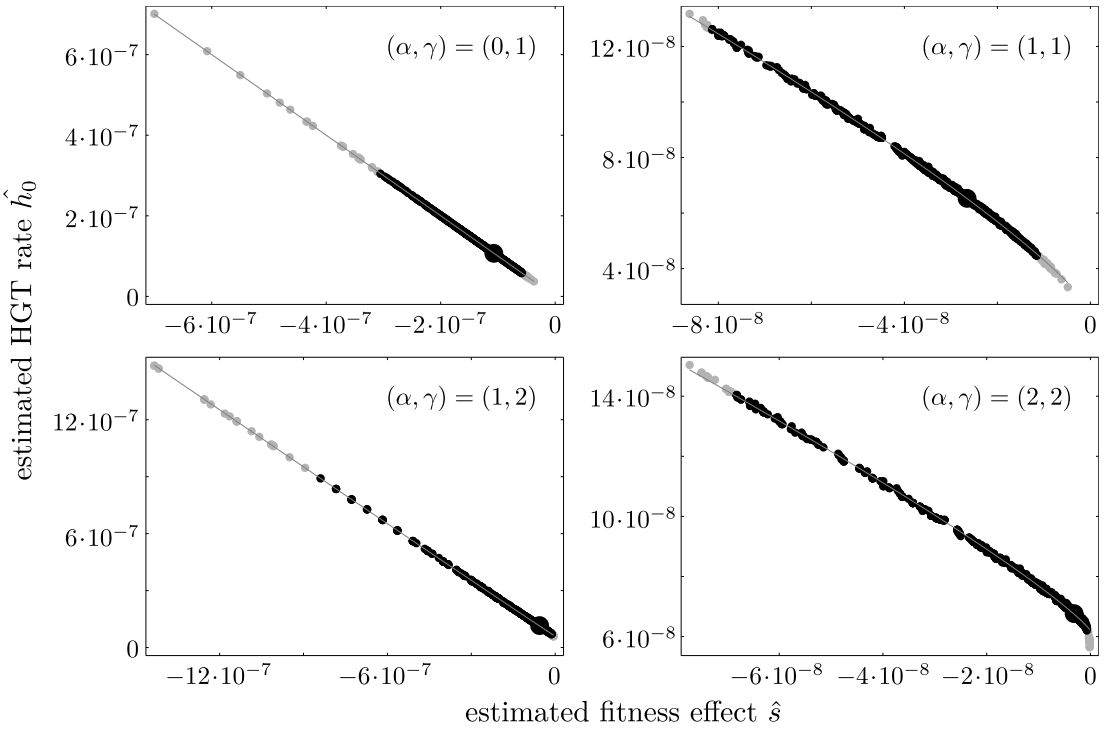

Figure 3.4: Bootstrapped pairs of $(\hat{s}, \hat{h_0})$ for $(\alpha, \gamma) \in \{(0,1),(1,1),(1,2),(2,2)\}$, based on 1000 resamplings of four computed IS count distributions. These four IS count distributions have been generated by using four estimate pairs $(\hat{s}, \hat{H})$, where each pair has been obtained from the IS*5* count distribution in a different random sample. Black dots lie inside and gray dots lie outside the 95% confidence interval of s. The original, estimated values of $(\hat{s}, \hat{h_0})$ are marked by large, black dots. The graphs of the shifted power function approximations $\hat{h_0} = a \cdot (-\hat{s})^b + c$ are shown as thin lines. The parameters (a, b, c) are $(0.998, 1.000, 9.8 \cdot 10^{-12})$ for $(\alpha,\gamma) = (0,1)$, $(0.065, 0.819, 2.5 \cdot 10^{-8})$ for $(\alpha,\gamma) = (1,1)$, $(1.113, 1.008, 6.0 \cdot 10^{-8})$ for $(\alpha,\gamma) = (1,2)$, and $(0.114, 0.860, 6.2 \cdot 10^{-8})$ for $(\alpha,\gamma) = (2,2)$.

3.4 Discussion

While an IS that provides a benefit to its host can rise to fixation through natural selection [Hall, 1999, Schneider and Lenski, 2004], the outcome of an infection with purely detrimental ISs is less clear. We have shown in an earlier paper that regardless of whether ISs are moderately beneficial or detrimental, the chances of a successful IS infection are small [Bichsel et al., 2010]. Here we are interested in the longer-term fate of an IS infection. Specifically, we investigate whether a purely detrimental IS can persist and reach the observed IS*5* count distribution in proteobacteria, where IS*5* mainly occurs. We are also interested in the fitness effect s and in the HGT rate h needed to reach this IS count distribution. We find the maximum likelihood estimates of s and h_0 by analysing 525 fully

sequenced genomes from 180 proteobacterial genera. We now discuss the main points of this study.

3.4.1 Purely detrimental ISs can persist if the HGT rate is larger than the fitness cost of an IS

The L-shaped IS*5* count distribution in 525 sequenced proteobacterial genomes (and presumably also in natural host cell populations) suggests that IS*5* (and probably all ISs with similar IS count distribution) is generally detrimental to its hosts. Our results support this suggestion and show that even purely detrimental ISs may persist and reach an IS count distribution similar to the one observed in IS*5* in sequenced genomes, provided that the HGT rate is larger than the fitness cost induced by one IS in the genome of an infected cell. This is in agreement with our earlier result based on a stochastic infection model [Bichsel et al., 2010]. The HGT rate in turn is larger than the fitness cost of one IS only if the possible increase in the infectiousness of a cell is smaller than the increase in the fitness cost with an increasing IS count. A small increase of the infectiousness with an increasing IS count is consistent with earlier observations that HGT is tightly regulated and depends on many different factors [Dröge et al., 1999]. If so, the influence of the IS count on the HGT rate, and therefore on the infectiousness of a cell, is probably small or even absent.

3.4.2 The observed IS*5* count distribution suggests that the replicative transposition rate of IS*5* is not down-regulated

Our model shows best agreement with the IS*5* count distribution in 797 of 1000 random samples, each containing 180 out of 525 proteobacterial genomes, if the replicative transposition rate increases linearly with the IS*5* count per genome, i.e. if replicative transposition is not regulated and copies of IS*5* transpose independently. This is in agreement with results published by Sawyer et al. [Sawyer et al., 1987]. Using branching processes to model the count distribution of several ISs in the ECOR collection of 71 natural isolates of *Escherichia coli*, these authors report that a linear dependence of the replicative transposition rate on the IS count agrees best with the available data in the collection for IS*5*. Sawyer et al. use for their analysis the ECOR collection, which is a smaller albeit more homogeneous

dataset than our collection of 525 sequenced proteobacterial genomes. Besides using a larger dataset, our analysis is based on an ordinary differential equation model that allows for interactions between cells and for density-dependent population growth and infection, which makes it more suitable to analyse the long-term fate of an IS infection than the branching process model used by Sawyer et al.

3.4.3 ISs might be effectively neutral to their hosts

Our model predicts a fitness effect in the range $\hat{s} \in [-10^{-7}, -10^{-9}]$ for IS*5* (see table 3.2). Considering that the effective population size of typical prokaryotes is of the order of $N_e \approx 10^8$ [Lynch, 2007, p. 92], IS*5* might therefore be effectively neutral or only slightly detrimental to its hosts. Hence, HGT is probably strong enough to enable IS*5* to persist and spread in a host cell population (see table 3.1). At the same time, our model predicts an unrealistically long time for IS*5* to approximately reach the final prevalence of infected cells, while the predicted time to approximately reach the IS*5* count distribution in infected cells only is much shorter. It therefore seems that the time scale of the infection process may be much larger than the time scale of the process that leads to an equilibrium distribution in the population of infected cells. While the former time scale is determined by the antagonistic actions of HGT and the fitness cost of one IS copy, the latter time scale is determined mainly by replicative transposition and the fitness cost of varying numbers of IS copies. This observation of different time scales leads us to suggest that IS*5* may have been at least occasionally and temporarily beneficial to its host cells, which can accelerate its spreading through single populations and through populations all over the world.

3.4.4 Caveats

The sequenced genomes stem from various proteobacterial cell populations all over the world and do not constitute a genome sample from a single population. At first sight, it is therefore not clear that we can compare the IS*5* count distribution in the sequenced genomes with the IS count distribution that our model of a single population predicts. However, we note that a very similar, L-shaped IS*5* count distribution has also been observed in the ECOR collection of 71 strains of *Escherichia coli* [Sawyer et al., 1987], which is a less heterogeneous

sample that covers a smaller taxonomic range than the proteobacterial genomes in our data set. This observation, together with the fact that the L-shaped IS count distribution can be observed in several other IS families [Sawyer et al., 1987, Wagner, 2006, Touchon and Rocha, 2007], motivates our assumption that this distribution does not depend on a specific IS and on the taxonomic scale. We thus assume that the same distribution does also exist in other ISs and on the smallest taxonomic scale, that of a cell population.

Another objection to our approach might be that we use IS*5* count data from phylogenetically related genomes to conduct a maximum likelihood analysis which assumes independence between observations. The genomes in our data set are related and their IS*5* counts are therefore not strictly independent of each other. Nevertheless, we have reduced this dependence by choosing only one genome per genus for the likelihood analysis. Furthermore, we generated 1000 sample data sets, each containing one genome per genus and repeated the maximum likelihood analysis for each of these data sets.

It might also be argued that the IS*5* count distribution in our data set is L-shaped because many ISs show certain DNA target specificities [Chandler and Mahillon, 2002]. IS*5* does in fact show some preference for the target sequence CTAG. However, because this nucleotide sequence is very short and therefore occurs frequently in host genomes, target specificity is probably not strong enough to limit the IS*5* count distribution noticeably in the IS count range on which we base our computations (0–60 copies of IS*5* per genome). This is supported by the observation that although most infected proteobacterial genomes have very low IS*5* counts, some genomes contain more than 40 copies of IS*5*. The same argumentation probably also holds for other ISs with some target specifities.

3.5 Appendix

3.5.1 Results for other replicative transposition and excision rates

We have repeated our calculations for another combination of the replicative transposition rate u and the excision rate e, this time at the lower end of the rate range described in table 3.1, namely $(u, e) = (10^{-9}, 10^{-11})$. Because the effect of the excision rate is small, and because the effect of an IS*5* infection on the normalized population density Z is again

negligible, the fitness effect s and the HGT rate h scale almost linearly with the assumed transposition rate u (see the differential equation system (3.2, 3.3)). This is exactly what can be observed. Table 3.3 shows for the four exponent pairs of (α, γ) that are most frequently not significantly worse than the best fitting pair in each sample the quartiles of the maximum likelihood estimates of s and of h for all choices of $\beta \in \{0, 1, 2\}$. We show only those HGT rates which lead to a stable equilibrium that can be reached by starting with a small proportion of infected cells (between 10^{-9} and 10^{-3}) carrying one copy of IS*5*.

(α, γ)	Quart.	$\hat{s}$	$\hat{h_0}$	$\hat{h_1}$	$\hat{h_2}$
$(0, 1)$	Q1	$-1.6 \cdot 10^{-9}$	$1.1 \cdot 10^{-9}$	–	–
	Q2	$-1.3 \cdot 10^{-9}$	$1.3 \cdot 10^{-9}$	–	–
	Q3	$-1.1 \cdot 10^{-9}$	$1.6 \cdot 10^{-9}$	–	–
$(1, 1)$	Q1	$-2.8 \cdot 10^{-10}$	$3.8 \cdot 10^{-10}$	$7.3 \cdot 10^{-11}$	–
	Q2	$-1.8 \cdot 10^{-10}$	$5.5 \cdot 10^{-10}$	$1.8 \cdot 10^{-10}$	–
	Q3	$-7.3 \cdot 10^{-11}$	$6.7 \cdot 10^{-10}$	$2.8 \cdot 10^{-10}$	–
$(1, 2)$	Q1	$-2.9 \cdot 10^{-10}$	$7.8 \cdot 10^{-10}$	$1.9 \cdot 10^{-10}$	–
	Q2	$-2.3 \cdot 10^{-10}$	$8.2 \cdot 10^{-10}$	$2.3 \cdot 10^{-10}$	–
	Q3	$-1.9 \cdot 10^{-10}$	$8.8 \cdot 10^{-10}$	$2.9 \cdot 10^{-10}$	–
$(2, 2)$	Q1	$-3.1 \cdot 10^{-11}$	$6.4 \cdot 10^{-10}$	$1.8 \cdot 10^{-10}$	$8.0 \cdot 10^{-12}$
	Q2	$-1.6 \cdot 10^{-11}$	$6.6 \cdot 10^{-10}$	$2.4 \cdot 10^{-10}$	$1.6 \cdot 10^{-11}$
	Q3	$-7.9 \cdot 10^{-12}$	$6.8 \cdot 10^{-10}$	$2.9 \cdot 10^{-10}$	$3.1 \cdot 10^{-11}$

Table 3.3: For a replicative transposition rate $u = 10^{-9}$ and an excision rate $e = 10^{-11}$, the table shows the four most frequent exponent pairs $(\alpha, \gamma) \in \{0, 1, 2\}^2$ that lead to model fits of the IS*5* count distribution that are not significantly worse than the best fit. For each pair (α, γ), the quartiles (Q1, Q2, Q3, where Q2 is the median) of the maximum likelihood estimates of the fitness effect s and of the HGT rate h_β for different scaling exponents $\beta \in \{0, 1, 2\}$ of the HGT rate are reported. Only HGT rates that lead to stable equilibria are shown. Observe that Q1 in $\hat{s}$ corresponds to Q3 in $\hat{h_\beta}$ and vice versa.

As can be seen when compared with table 3.2 in the main text, the quartiles of the maximum likelihood estimates of s and h scale almost perfectly linearly with the new choice for the replicative transposition rate u.

We draw the same conclusions as in the main text: IS*5* seems to be effectively neutral (even more so for this parameter combination of u and e), and HGT ist most probably strong enough to overcome the fitness cost caused by a copy of IS*5* in the host cell genome.

3.5.2 Population dynamics of an IS infection in dependence of the model parameter set

Figure 3.5 shows the computed population dynamics of a host cell population that has been infected with a fraction of 10^{-6} cells harboring one IS copy in their genomes. We chose $r = 1$, $u = 10^{-7}$, $e = 10^{-10}$, and the maximum likelihood estimates $\hat{s} = -1.1 \cdot 10^{-7}$ and $\hat{h_0} = 1.1 \cdot 10^{-7}$ for the exponent pair $(\alpha, \gamma) = (0, 1)$ to compute the infection dynamics based on the equation system 3.1.

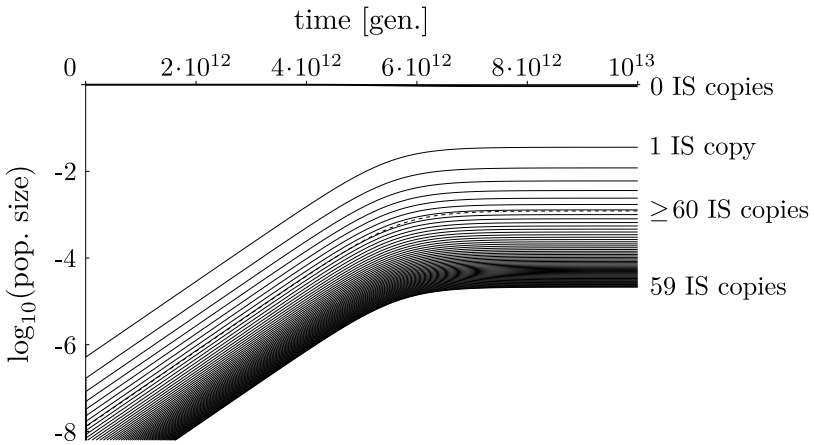

Figure 3.5: Computed population dynamics of a host cell population infected with a fraction of 10^{-6} cells harboring one IS copy in their genomes. We chose $r = 1$, $u = 10^{-7}$, $e = 10^{-10}$, $(\alpha, \gamma, \beta) = (0, 1, 0)$, and the corresponding maximum likelihood estimates $\hat{s} = -1.1 \cdot 10^{-7}$ and $\hat{h_0} = 1.1 \cdot 10^{-7}$ as model parameters. The curves for cells harboring different numbers of IS copies in their genomes are indicated on the right. The curve for cells harboring 0 IS copies in their genomes is shown in bold, and the curve for cells harboring at least 60 IS copies in their genomes is shown as a dashed line. Time is measured in cell generations. Note the logarithmic scale on the vertical axis.

Observe that the IS count distribution at 10^{13} generations in figure 3.5 is the same as the computed IS count equilibrium distribution in figure 3.2. As can be seen in figure 3.5, on the one hand, it takes a very long time to reach the population equilibrium of uninfected and infected cells. On the other hand, it takes a much shorter time to reach an equilibrium in the IS count distribution among infected cells only.

To illustrate the influence of different model parameters on the population dynamics and on the final IS count distribution, figure 3.6 shows the population dynamics if each of the parameters u, e, s, and h_0 has been separately set to one tenth of its original value as used in figure 3.5.

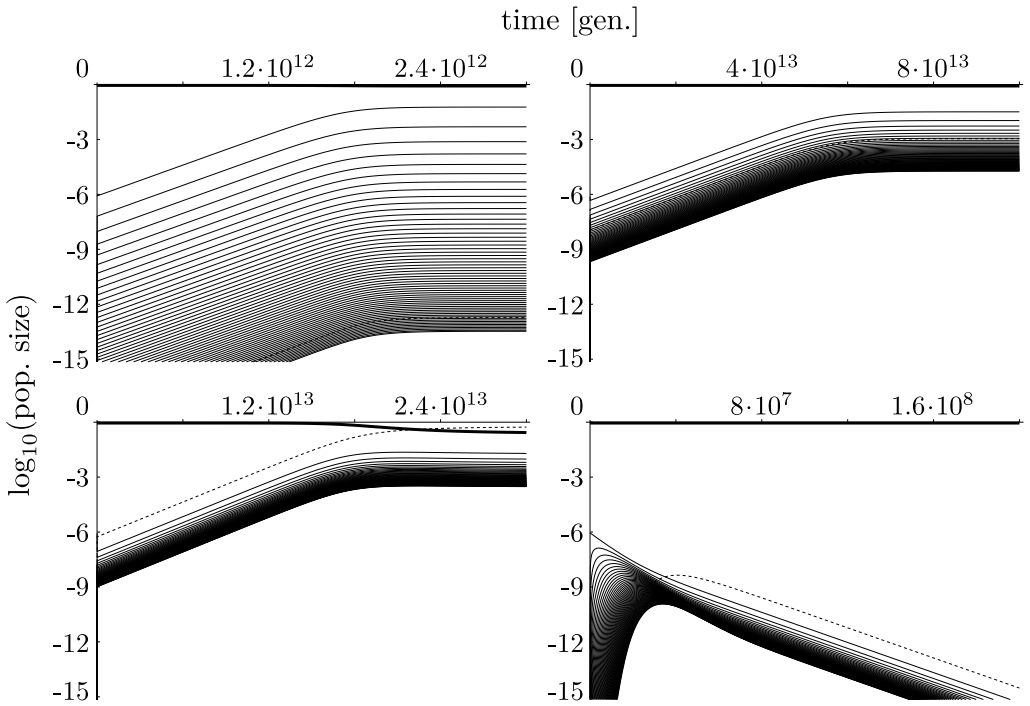

Figure 3.6: Computed population dynamics of a host cell population infected with a fraction of 10^{-6} cells harboring one IS copy in their genomes, for different parameter sets. The original model parameters are the same as in figure 3.5: $r = 1$, $u = 10^{-7}$, $e = 10^{-10}$, $(\alpha, \gamma, \beta) = (0, 1, 0)$, $s = -1.1 \cdot 10^{-7}$, and $h_0 = 1.1 \cdot 10^{-7}$. For each of the four graphs, exactly one parameter has been changed compared to the original parameter set: $u = 10^{-8}$ (top left), $e = 10^{-11}$ (top right), $s = -1.1 \cdot 10^{-8}$ (bottom left), and $h_0 = 1.1 \cdot 10^{-8}$ (bottom right). In each graph, the curve for cells harboring 0 IS copies in their genomes is shown in bold, and the curve for cells harboring at least 60 IS copies in their genomes is shown as a dashed line. Time is measured in cell generations. Note the logarithmic scale on the vertical axis.

Compared with figure 3.5, figure 3.6 shows that reducing the transposition rate leads to a steeper final IS count distribution, with less cells harboring high numbers of IS copies in their genome (top left graph with $u = 10^{-8}$). Figure 3.6 also shows that reducing the IS excision rate does not change the final IS count distribution noticeably, but it takes a longer time to reach this final distribution (top right graph with $e = 10^{-11}$). Reducing the fitness cost tenfold leads to a population dominated by infected cells with the highest IS count allowed in our model, noticeably reducing the normalized density of uninfected cells (bottom left graph with $s = -1.1 \cdot 10^{-8}$). Reducing the HGT rate below the fitness cost, in turn, does not allow the population of infected cells to persist (bottom right graph with $h_0 = 1.1 \cdot 10^{-8}$).

4. The Dynamics of an IS Infection in a Spatially Structured Environment

Manuel Bichsel, Andrew D. Barbour, Andreas Wagner; submitted to *Journal of Theoretical Biology*

Abstract

Bacterial insertion sequences, the simplest form of autonomous mobile DNA, depend on their prokaryote hosts to spread in a spatially structured environment. We use a spatially explicit metapopulation model to simulate the spread of an insertion sequence that can have both detrimental and beneficial effects on its host cell. We find that, on the one hand, the spatial structure of the metapopulation and cell dispersal between subpopulations have no strong effect on the time to full infection of the metapopulation. On the other hand, factors that influence the IS infection dynamics within a subpopulation have a strong effect on the time to full infection of the metapopulation. These factors include the fitness benefit of an insertion sequence and the rate of horizontal gene transfer. We also find that the infection process of a metapopulation is very erratic in its early phase, and that the infection's success depends critically on the initially infected subpopulation. Finally, we show that even though beneficially infected cells speed up the infection process, they need not necessarily constitute the majority of infected cells.

4.1 Introduction

Mobile DNA has been fascinating researchers since its discovery in the 1940s by Barbara McClintock [McClintock, 1950]. Why does it persist, even though its effects are detrimental

to its host cells on average? The persistence of mobile DNA is especially puzzling in prokaryotes. While even detrimental mobile DNA may spread in a sexually reproducing eukaryote, especially if the mobile DNA's effects are recessive [Hickey, 1982, Charlesworth et al., 1994, Brookfield, 2005], the detrimental effects of mobile DNA cannot be masked in this way in an asexually reproducing prokaryote. In addition, due to the generally high effective population size of prokaryotes, even small detrimental fitness effects of mobile DNA may cause strong negative selection. The spread of mobile DNA in prokaryotes is thus more difficult to explain.

In this paper, we study the spread of insertion sequences (ISs), the simplest form of autonomous mobile DNA, in spatially structured metapopulations of prokaryotes. ISs are short DNA sequences (0.7–2.5 kb) that typically encode only one enzyme, transposase, which enables transposition. During transposition, an IS usually gets excised and inserted into another location in the genome (conservative transposition), but occasionally the IS is copied during the transposition process (replicative transposition) [Chandler and Mahillon, 2002]. While the number of active IS copies in a genome increases through replicative transposition, it decreases through IS excision and mutations that render transposase ineffective. Insertion sequences are vertically transmitted through inheritance, but ISs can also be horizontally transmitted by horizontal gene transfer (HGT) between prokaryotes.

ISs in general have a detrimental fitness effect on their host cell, not only due to their transposition activity, but also because of genome rearrangements that can occur if a genome contains more than one IS copy [Galas and Chandler, 1989, Kleckner, 1989, Schneider and Lenski, 2004]. Occasionally, though, ISs benefit their hosts. Many ISs contain an outwardly directed, entire or partial promoter that can increase the expression of a nearby gene [Galas and Chandler, 1989]. Furthermore, two synchronously transposing, flanking ISs can mobilize intermediary genes. These genes often confer resistance to antibiotics, encode toxins, or allow for new metabolic functions [So and McCarthy, 1980, Berg, 1989, Top and Springael, 2003]. The composite transposon (also called compound transposon) that has thus been created can then insert into a plasmid and spread through a host cell population. A still unresolved question is whether ISs persist because they are occasionally beneficial to their hosts [Blot, 1994, Shapiro, 1999, Schneider and Lenski, 2004] or because HGT is strong

enough to overcome their detrimental fitness effects [Dawkins, 1976, Doolittle and Sapienza, 1980, Orgel and Crick, 1980, Charlesworth et al., 1994, Nuzhdin, 1999]. In earlier papers, we showed that a purely detrimental IS infection can successfully invade an uninfected host cell population, provided that the HGT rate exceeds the detrimental effect of ISs on a host cell [Bichsel et al., 2010]. For a specific IS family, we also estimated the HGT rate that would be needed to reach the distribution of IS copies per genome which can be observed in the wild. We showed that this HGT rate is well within the range of HGT rates estimated in the wild, but that the infection process would take an unrealistically long time if it depended only on HGT [Bichsel et al., 2013]. We then concluded that beneficial effects of an IS infection on infected cells, even though they may be temporary, probably play an important role in speeding up the infection process. This is in accordance to an earlier finding of one of us [Wagner, 2006] who has shown that the sequence divergence of IS copies within genomes is much lower than between genomes, indicating that ISs might undergo 'burst and bust' cycles of infection and extinction in local populations.

The fact that an IS infection usually occurs in a spatially structured environment may also play a role in the infection's dynamics. Although theoretical predictions about IS infection dynamics in a single population without spatial structure have been made (some of them by us) [Sawyer and Hartl, 1986, Basten and Moody, 1991, Bichsel et al., 2010], there exists to our knowledge no analysis of IS infection dynamics in a spatially structured environment. In such an environment, ISs may spread to new prokaryotic host populations through the dispersal of infected cells. While the Baas Becking hypothesis for microorganisms, "*Everything is everywhere, but the environment selects*" [Baas Becking, 1934, transl.], is still debated among microbiologists (for a review, see [Fierer, 2008]), it has been shown that microorganisms, especially spore-forming prokaryotes, can spread not only through migrating host animals [McCallum et al., 2003], but also directly through the air by wind [Burrows et al., 2009, Kellog and Griffin, 2006] or through the oceans by water currents [Mao and Grogan, 2012, Freel et al., 2012] over long (even intercontinental) distances. Many prokaryote species thus seem to have a global distribution, although other prokaryote species, e.g. the extremophiles, seem to show a more restricted, local distribution [Papke and Ward, 2004]. However, even in a globally distributed prokaryote host species, the spreading of an

IS does not occur instantaneously, and IS infection dynamics may be affected by the spatial structure of the environment.

A large literature exists about spatial invasion processes in ecology [Tilman and Kareiva, 1997, Hanski, 1999] and about spatial infection processes in epidemiology [Keeling and Rohani, 2008b]. Some phenomena that may be observed during invasion or infection processes in the wild are directly linked to space. One such phenomenon is the appearance of infection waves, where the geographical area containing infected individuals or subpopulations shows a well-defined, expanding front line. Another relevant phenomenon is the extinction/rescue effect, whereby an invading species may become temporarily extinct in a specific habitat patch (subpopulation), only to be rescued by immigrating individuals from another patch. In its extreme form, this effect can lead to a source-sink dynamics, where an invading species only persists in a specific patch because of constant immigration from other patches [Hanski, 1999]. Phenomena like these can only be analyzed with spatially explicit models, where subpopulations or even individuals occupy specific spatial locations. With such a model, one can then explore the effect of different spatial distributions of subpopulations on invasion or infection speeds, and examine pattern-formation during invasion or infection.

Here we use a metapopulation model to simulate the spreading of an IS infection. The metapopulation consists of spatially separate subpopulations, where all but one of the subpopulations initially contain only uninfected cells. The infection dynamics of ISs in the metapopulation is determined by *local* processes within each subpopulation, such as HGT and the competition between infected and uninfected host cells, and by *global* processes between subpopulations, such as cell dispersal. We then use this model to address the following three main questions concerning beneficial and detrimental IS infections in a spatially structured environment:

1. How do spatiality and local or global processes influence IS infection speed?
2. What is the role of the initially infected subpopulation in the infection process?
3. How large is the fraction of beneficially infected cells among all infected cells as a function of time in infected subpopulations?

4.2 Model and Methods

The model we used in our simulations consists of two levels: a subpopulation level (figure 4.1A) and a metapopulation level (figures 4.1B–D).

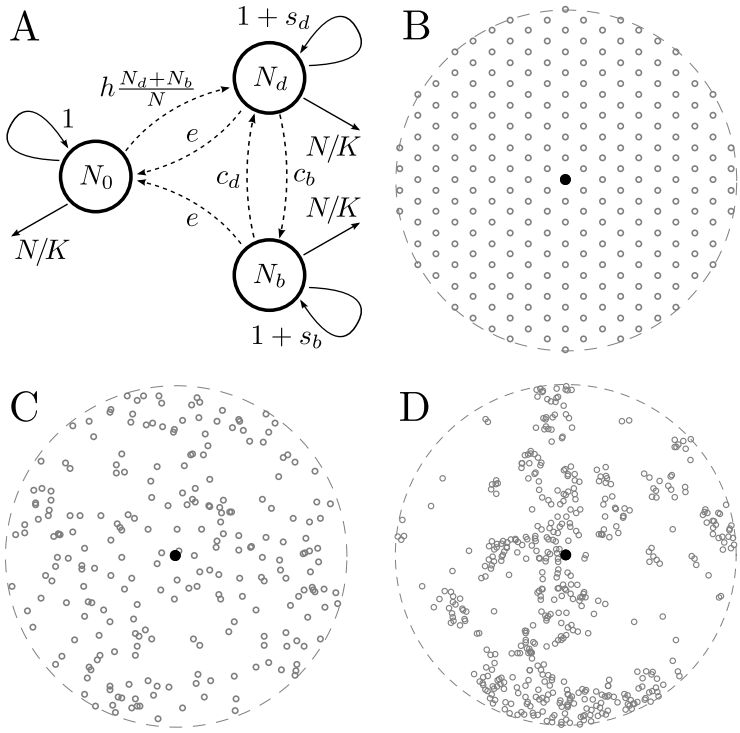

Figure 4.1: *Model design.* A: Subpopulation level; K is the carrying capacity; N_0, N_d and N_b are the cell counts of uninfected, detrimentally infected and beneficially infected cells, respectively, and $N = N_0 + N_d + N_b$ is the total cell count; $s_d < 0$ and $s_b > 0$ indicate the fitness cost of a detrimental infection and the fitness benefit of a beneficial infection, respectively; c_d and c_b are the conversion rates to detrimental and beneficial infection, respectively; e is the IS excision rate; h is the HGT rate. All rates are per cell and cell generation.
B–D: Metapopulation level; the regular (B), uniform random (C) and clustered (D) spatial distributions of 241 subpopulations inside a circular region with a radius of 100 km are shown; the initially infected subpopulation in the center is indicated as a black closed circle; the landscape in the clustered spatial distribution of subpopulations (D) has a contagion index of 0.4.

On the subpopulation level, we assume that each host cell can carry at most one single IS in its genome to keep the simulations feasible. This is not a strong limitation, as the IS count distribution of all IS families in the wild is strongly L-shaped, i.e. for any IS family, most genomes contain no IS copy, many genomes contain one copy, and only few contain more than five copies [Wagner, 2006, Touchon and Rocha, 2007, Bichsel et al., 2013]. To

reduce model complexity, and to increase simulation speed, it is therefore reasonable to allow for at most one IS per genome. In our model, an IS insertion can either have a detrimental or a beneficial effect on host cell fitness. Empirical data show that the fitness effect of an IS depends on the location of the insertion in the genome. In a few locations, an IS may have a beneficial fitness effect, for example by promoting the expression of a nearby gene [Galas and Chandler, 1989]. At most other locations, the same IS may have a detrimental fitness effect, for example by inserting into a gene, thus silencing the gene [Rodriguez et al., 1992]. Noncoding regions, into which an IS could safely insert, constitute only a small fraction of about 10% of prokaryotic genomes [Lynch, 2007], and even noncoding regions include many regulatory DNA sequences that are sensitive to disruption. Genome locations with beneficial side effects are therefore rare in comparison with genome locations with detrimental side effects. With these observations in mind, we assume that the conversion from a detrimentally infected cell to a beneficially infected cell by conservative transposition occurs with a rate c_b that is 1000 times lower than the rate of conversion c_d from a beneficially infected cell to a detrimentally infected cell. Because conservative transposition is the only way to switch between beneficial and detrimental infection in our model, and because transposition into a genome location with beneficial side effects is very rare, we set the conversion rate c_d of beneficial to detrimental infection to the conservative transposition rate in the wild. An IS can get excised and lost (or be disabled) with rate e [Kleckner, 1989, Sousa et al., 2013], and an IS can be copied from an infected host cell's genome to an uninfected host cell's genome by HGT with rate h [Frost et al., 2005]. Because genome locations that lead to a beneficial infection are rare, we assume that all cells that acquire an IS by HGT will initially be detrimentally infected. Transposition, excision and HGT rates reported in the literature cover a wide range of values (see appendix). Besides using a default parameter set, with which we conducted most of our simulations, we therefore also performed simulations where we systematically varied the parameters of local processes over a wide range of values. Table 4.1 shows for each local parameter its default value and the parameter range that we explored.

The effective size N_e of prokaryote populations usually exceeds 10^8 individuals [Lynch and Conery, 2003]. We therefore assume that each well-mixed subpopulation has a carrying

Symbol	Description	Default value	Range of values
s_d	detrimental fitness effect	-10^{-4}	$[-10^{-3}, -10^{-5}]$
s_b	beneficial fitness effect	10^{-4}	$[10^{-7}, 10^{-1}]$
c_d	detrimental conversion rate	10^{-6}	$[0, 10^{-6}]$
c_b	beneficial conversion rate	10^{-9}	$[0, 10^{-9}]$
e	excision rate	10^{-9}	$[0, 10^{-9}]$
h	HGT rate	10^{-4}	$[10^{-7}, 10^{-1}]$

Table 4.1: Default values and ranges of values of local parameters used in our simulation model. All rates are given as numbers of events per cell and generation.

capacity of $K = 10^9$ host cells (computational limitations prevented us from exploring much larger populations). At the start of the simulation, all subpopulations contain 10^9 uninfected cells. One subpopulation – the initially infected subpopulation – contains an additional 100 beneficially infected cells. We chose an initial number of 100 infected cells instead of only one infected cell as a compromise between two conflicting requirements. On the one hand, as the number of initially infected cells increases, so does the probability that a metapopulation becomes fully infected during simulation. Because we are mainly interested in metapopulations that become fully infected, we need fewer simulations if we increase the number of initially infected cells. On the other hand, an increasing number of initially infected cells decreases the time for a metapopulation to become fully infected. Increasing the number of initially infected cells thus leads to unrealistically short times for metapopulations to become fully infected. In exploratory simulations starting with one beneficially infected cell, the median time needed to reach 100 infected cells was 60 generations, with an interquartile range of 45 generations. This is very small compared with the time to full infection of a metapopulation, which typically is of the order of $2 \cdot 10^5$ generations. It is therefore reasonable to choose an initial number of 100 beneficially infected cells.

We simulated the IS infection dynamics in each subpopulation by using the tau-leaping algorithm [Gillespie, 2001, Cao et al., 2006]. We used this algorithm to calculate the length of each time step in the simulation and to determine the numbers N_0 of uninfected cells, N_d of detrimentally infected cells, and N_b of beneficially infected cells at the end of each time step.

On the metapopulation level, we simulated cell dispersal by exchanging cells between all

subpopulations every 10 generations. We computed the number of cells that migrate from one subpopulation to another based on a Poisson distribution. The distribution's mean is computed using a cell dispersal rate function that relates the mean number of cells migrating from one subpopulation to another during one generation to the distance (in kilometers) between the two subpopulations. We based our default rate function on data that two authors [Roberts and Cohan, 1995] obtained by applying a cladistic method [Slatkin and Maddison, 1989] on *Bacillus subtilis* and *Bacillus mojavensis* nucleotide sequences from three continents. Briefly, in this method, one treats the geographic location of each sequence as a multistate character and adds the location to the sequence phylogeny. Then one derives the minimal number of migration events that are necessary to obtain the observed distribution of multistate characters over the phylogeny. From this number of migration events, the average number of migrating individuals between subpopulations per generation is calculated. We fitted an inverse power law function to the data points of cell dispersal rate versus distance, which leads to a better fit (residual sum of squares $RSS = 127.3$) than other functions, e.g. a negative exponential function ($RSS = 1145.5$). This is in agreement with observations of long-distance dispersal of pollen and plant seeds, where the dispersal kernel usually has a fat tail, i.e. the tail drops off slower than in an exponential function [Nathan et al., 2008, Shaw et al., 2006, Bullock and Clarke, 2000]. This best-fitting inverse power law function is our default rate function. It relates the dispersal rate r (in cells per generation) to the distance d (in kilometers) and has the form $r = 239.6 \cdot (d + 0.1)^{-0.53}$. We added the constant 0.1 to the distance to avoid a singularity at $d = 0$. In view of the uncertainty about dispersal rates over long distances, we also conducted simulations using rate functions with different proportions of dispersal over short and long distances, while keeping the same mean dispersal rates in a metapopulation with regularly distributed subpopulations on the vertices of a hexagonal grid (see figure 4.1B), to allow comparisons between rate functions. To explore the effects of steeper inverse power law functions with exponents of at least -2 that have been estimated by other authors for pollen and plant seeds [Bullock and Clarke, 2000, Shaw et al., 2006], we used an exponent of $4 \cdot (-0.53) = -2.13$ and readjusted the multiplicative constant so that the mean dispersal rate in a metapopulation with regularly distributed subpopulations stays the same as for the default rate function.

This led to the rate function $r = 91\,728.4 \cdot (d + 0.1)^{-2.13}$. We also included two extreme dispersal rate functions: a constant function, where the dispersal rate does not depend on the distance between two subpopulations, and a nearest neighbour function, where cell dispersal occurs only between immediately neighbouring subpopulations, i.e. there exists a threshold distance below which dispersal is constant and above which dispersal is impossible. We again chose the constants so that the mean dispersal rate in a metapopulation with regularly distributed subpopulations is the same as for the default rate function. This led to the constant rate function $r = 24.5$ for all subpopulations. For the nearest neighbour rate function, the cell dispersal rate to the nearest six subpopulations on the hexagonal grid is $r = 1065.3$ and $r = 0$ to all other subpopulations.

In our model, all subpopulations are spatially distributed inside a circular region. We chose a radius of 100 km for this region, because the region is then large enough to permit spatial phenomena like infection waves to take place, while being small enough to keep the number of subpopulations low and allow for manageable simulation times. A metapopulation can be thought of as a small geographical (eco)region with a spatially distributed habitat suitable for host cells [Olson et al., 2001].

We conducted most of our simulations with the 241 subpopulations that lie inside the circular region mentioned above on the vertices of a hexagonal lattice with an edge length of 12.5 km (see figure 4.1B). We used exploratory simulations with halved and doubled edge lengths to ensure that our choice of edge length would not influence our results. In order to assess the effect of different spatial distributions of subpopulations, we also used two more types of spatial distribution. In the first, the 241 subpopulations show a uniform random distribution (see figure 4.1C). In the second, the 241 subpopulations show a clustered distribution, determined by a mixture of habitable and uninhabitable landscape types with a contagion index of 0.4 (see figure 4.1D). The contagion index [O'Neill et al., 1988,Li and Reynolds, 1993] is a measure of the clumpiness of a landscape, where a value of zero signifies no clumping, i.e. a fine-grained mixture of landscape types, and a value of one signifies that the landscape consists of only one type. For comparison, we also conducted simulations with a spatially unstructured single population with the same initial number of cells as the spatially structured metapopulations ($241 \cdot 10^9$ uninfected cells plus 100

beneficially infected cells).

We use the time to full infection as a simple indicator of the infection dynamics. We define the time to full infection as the time until at least 99% of all subpopulations have an infection prevalence of at least 95%. This definition ensures that isolated subpopulations with low cell dispersal rates will not unduly distort our results. For the large, spatially unstructured population with a carrying capacity of $241 \cdot 10^9$ cells, we correspondingly define the time to full infection as the time to reach an infection prevalence of at least $0.95 \cdot 0.99 = 94.05\%$. Since the time to full infection usually has a skewed distribution, we use the nonparametric Wilcoxon-Mann-Whitney test with a significance level of $\alpha = 0.05$ to compare times between different metapopulations.

We wrote the simulation program in C++ with gcc (version 4.6.4), using the Boost libraries (version 1.49.0). For metapopulations with subpopulations on a hexagonal lattice, we conducted 5000 simulations per metapopulation, of which typically about 50 simulations led to full infection. For metapopulations with randomly distributed subpopulations, we conducted 100 simulations on each of 50 different realisations of a random subpopulation distribution. All times are given in cell generations. The time resolution of the simulation output is 500 generations, i.e. every 500 generations, we report the number of uninfected, detrimentally infected and beneficially infected cells existing in every subpopulation. In addition, we also report the number of immigrating or emigrating cells for every subpopulation during the previous 500 generations. We analyzed all data using Mathematica (versions 8 and 10).

4.3 Results

4.3.1 Early on, an IS infection is an erratic process

We first analysed the early phase of an IS infection, when infected cells are restricted primarily to the initially infected subpopulation. Whether an infection of the metapopulation will succeed or not depends mainly on the fate of the infection in the initially infected subpopulation during this early phase, because it is unlikely that any of the few infected cells will be dispersed to other, still uninfected subpopulations. Using default values for the

local (see table 4.1) and global parameters of our metapopulation, with subpopulations on a hexagonal grid (see figure 4.1B), only 49 out of 5000 simulations led to full infection of the metapopulation. Moreover, only in one out of these 49 full infections of the metapopulation did the infection in the initially infected subpopulation die out temporarily and was rescued by another subpopulation which it had infected before. This last observation illustrates the importance of the initially infected subpopulation for the infection process of the metapopulation. Due to the low number of infected cells in the initially infected subpopulation, the early phase is dominated by stochastic effects, as we have already shown in an earlier paper [Bichsel et al., 2010].

The low number of 49 full infections of the metapopulation in 5000 simulations with default parameters is to be expected, as we show with the following calculation. The absolute value of the fitness cost $s_d = -10^{-4}$ equals the HGT rate h, so that the probability of an infection to persist depends almost exclusively on the fitness benefit $s_b = 10^{-4}$ of a beneficial infection, as HGT is only just strong enough to compensate for the fitness cost of a detrimental infection [Bichsel et al., 2010]. Furthermore, the IS conversion rates c_d and c_b are very low compared with the fitness benefit of a beneficial IS infection. The probability of an infection starting with a single, beneficially infected cell to persist and ultimately spread through the subpopulation and the metapopulation therefore corresponds to the probability of a beneficial mutant gene to spread through the sub- and the metapopulation. According to a result by Haldane [Haldane, 1927], a dominant mutant gene with a small selective advantage s, so that the expected number of offspring is $1 + s$, has a fixation probability of $2s$ in a random mating population. Only considering the small fitness benefit s_b of an IS but not HGT and IS conversion, the expected number of offspring of a beneficially infected cell is $2 \cdot (b + s_b) \,/\, (b + s_b + d) \approx 1 + s_b/2$, where $b = d = 1$ are the birth and death rates of an uninfected cell in a subpopulation at carrying capacity. In our model, the selective advantage assumed by Haldane is therefore $s = s_b/2$, and the probability of a beneficially infected cell to persist and first spread through the subpopulation and then through the metapopulation is $2s = s_b = 10^{-4}$. As the 100 initially infected cells act independently of each other during the early phase of a subpopulation infection, when their number is low compared to the total number of host cells in the subpopulation, the number of persisting

cell lineages, each starting with one beneficially infected cell, has a binomial distribution $\mathrm{Bi}(n;p)$ with $n = 100$ and $p = s_b = 10^{-4}$, and the probability of an infection with 100 beneficially infected cells to persist is therefore $1-(1-s_b)^{100} = 1-(1-10^{-4})^{100} \approx 0.01$. We thus expect about 50 out of 5000 simulations to lead to full infection of the metapopulation.

We call a subpopulation successfully infected if the number of infected cells in the subpopulation increases faster than the number of immigrating infected cells, i.e. if the infection in the subpopulation can spread on its own, without any immigration of infected cells. The median time for the initially infected subpopulation to successfully infect a second subpopulation is $2.0 \cdot 10^4$ generations, but with a large interquartile range of $1.8 \cdot 10^4$ generations.

We also find that a subpopulation usually gets (re)infected several times before it is successfully infected, i.e. most infections of a subpopulation die out. For example, a median number of 57.5 unsuccessful infection attempts happen in the first subpopulation that gets infected by the initially infected subpopulation until that first subpopulation gets successfully infected. This number is actually a lower limit for the number of unsuccessful infection attempts, because our simulations record the number of infected cells at fixed time points, so that unsuccessful infections between those time points may be missed.

Taken together, these observations demonstrate that the infection process is very erratic during its early phase, when both the number of infected subpopulations and the number of infected cells per subpopulation are low. Over time, when a larger fraction of cells in a subpopulation gets infected, the infection process becomes more deterministic. For example, the coefficient of variation of the time to reach an infection prevalence of 1% in the initially infected subpopulation (starting with a prevalence of 10^{-7}) is 15.0%, while the coefficient of variation of the time that passes between reaching an infection prevalence of 94% and of 95% in the initially infected subpopulation is only 5.7%.

4.3.2 An IS infection is not strongly slowed down by spatiality

We next examined how the spatial distribution of subpopulations in a metapopulation influences the IS infection dynamics. Figure 4.2 shows the time to full infection for different spatial distributions, using default values for the local parameters of all metapopulations.

The nonspatial metapopulation consists of a well-mixed single population starting with the same number of cells as the other metapopulations. The median time to full infection for this large, single population is $1.82 \cdot 10^5$ generations. Although the difference in time between the nonspatial population and the spatially structured metapopulations is statistically significant ($p < 0.001$ in pairwise Wilcoxon-Mann-Whitney tests), the median times to full infection for the three metapopulations with a spatial structure (median times between $1.97 \cdot 10^5$ and $2.01 \cdot 10^5$ generations) are at most 10.7% higher than the median time for the nonspatial population.

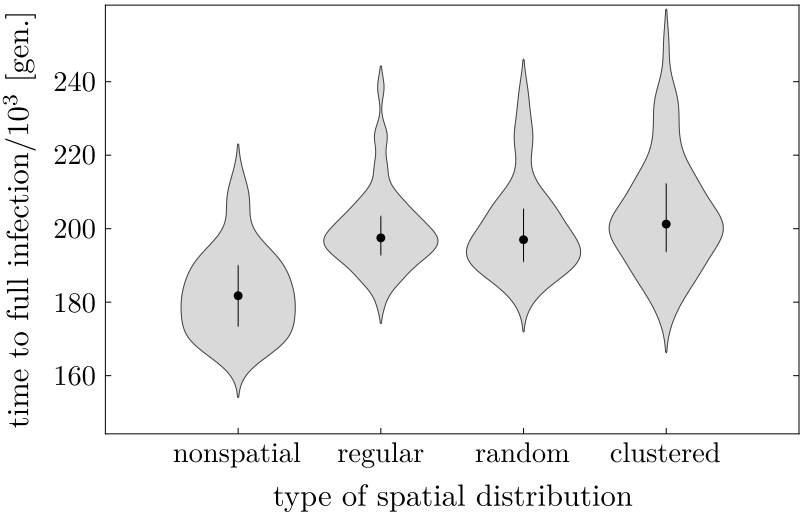

Figure 4.2: *The spatial distribution of subpopulations influences IS infection dynamics only weakly.* Violin plot showing a kernel density estimate (shaded region), the median (dot) and the first and third quartile (endpoint of whiskers) of the time to full infection for a nonspatial, single population and the following three spatially structured metapopulations: regular (subpopulations on a hexagonal lattice), random (subpopulations uniformly distributed), clustered (subpopulations uniformly distributed in a clustered landscape). Number of observations (left to right): 62, 49, 57, and 56 simulations that led to fully infected metapopulations (out of 5000 simulations).

Figure 4.2 also shows that the type of spatial distribution (regular, random, or clustered) does not noticeably influence the median time to full infection ($p \geq 0.097$ in pairwise Wilcoxon-Mann-Whitney tests). This may be a consequence of the fact that the default power law cell dispersal rate function, which is based on dispersal data from the wild, is relatively flat. For this function, the dispersal rate only drops from 70.3 cells per generation at a distance of 10 km between subpopulations to 20.9 cells per generation at a distance of 100 km between subpopulations. With this rate function, strongly varying distances

between a subpopulation and its nearest neighbours, which are a characteristic of clustered landscapes, do therefore not pose any difficulties for the IS infection and will not slow it down noticeably.

4.3.3 The shape of the dispersal function has only a limited influence on the infection speed

We also explored the effect of the cell dispersal rate function on the IS infection dynamics. Figure 4.3 shows the time to full infection for four different cell dispersal rate functions: constant, default power law, steep power law, and nearest neighbour (see 'Model and Methods').

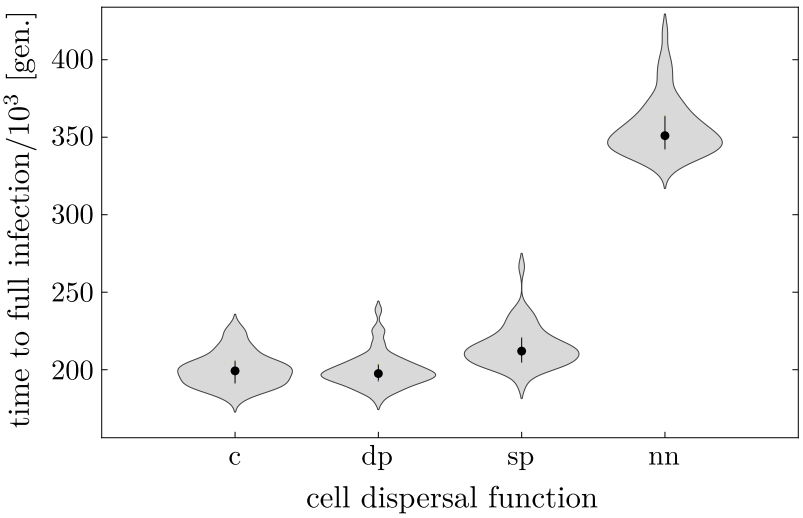

Figure 4.3: *Realistic, data-based cell dispersal influences IS infection dynamics only weakly.* Violin plot showing a kernel density estimate (shaded region), the median (dot) and the first and third quartile (endpoint of whiskers) of the time to full infection for the constant (c), default power law (dp), steep power law (sp) and nearest neighbour (nn) cell dispersal function. Number of observations (left to right): 58, 49, 48, and 51 simulations that led to fully infected metapopulations (out of 5000 simulations).

The times to full infection for the constant rate function (median time: $1.99 \cdot 10^5$ generations) and the default power law rate function (median time: $1.98 \cdot 10^5$ generations) are not significantly different ($p = 0.7$). In contrast, the times to full infection for both power law rate functions (median time: $1.98 \cdot 10^5$ and $2.12 \cdot 10^5$ generations for the default and the steep power law rate function, respectively) and the nearest neighbour rate function (median time: $3.51 \cdot 10^5$ generations) are all significantly different from each other (pairwise

tests, $p \leq 0.013$). However, the difference in the median time to full infection between the two power law rate functions is small (7.3%), and only the nearest neighbour dispersal rate function leads to a substantially longer median time to full infection than the other functions (e.g. 77.7% longer than for the default power law rate function). Figure 4.3 shows that the times to full infection for the default power law rate function, which is based on dispersal data from the wild, and two rate functions with flatter (constant) or steeper (steep power law) shape are very similar. Even for the nearest neighbour rate function, the time to full infection has the same order of magnitude. For quite different shapes of the dispersal rate function, the infection speed does therefore not vary strongly inside a metapopulation with regional extent.

We examined the effect of changing the total cell dispersal rate while keeping the overall shape of the dispersal function the same. To this end, we used the default power law dispersal function $r(d) = c \cdot (d + 0.1)^{-0.53}$, where $c = 239.6$ (see 'Model and Methods'), and decreased or increased the multiplicative constant c by up to two orders of magnitude. Based on 5000 simulations for ten- and hundredfold smaller or larger values of c, we observed that such a decrease/increase of the dispersal rate leads to significantly longer/shorter times to full infection than those obtained with the default rates (pairwise tests, $p < 0.001$). Specifically, while a tenfold reduced dispersal rate leads to a substantially increased time to full infection (+22.2%), a tenfold increased dispersal rate leads to a much smaller decrease in the time to full infection, from $1.98 \cdot 10^5$ to $1.84 \cdot 10^5$ generations (-7.1%). This is not unexpected, because the time to full infection for a single, spatially unstructured population with carrying capacity $K = 241 \cdot 10^9$ is already as high as $1.82 \cdot 10^5$ generations, and there is not much room for a decrease in the time to full infection. In summary, while the shape of the dispersal function, i.e. different proportions of dispersal over short and long distances, does not strongly influence the time to full infection, the total amount of dispersed cells has a strong influence on this time.

4.3.4 Both HGT rate and fitness benefit of an IS strongly influence infection speed

We next examined how the fitness benefit s_b of an IS and the HGT rate h influence the infection dynamics. In contrast to the spatial distribution of subpopulations and the cell dispersal rate, these two parameters both belong to local processes within a subpopulation.

We first set the HGT rate to its default value $h = 10^{-4}$. Figure 4.4 shows that the time to full infection strongly depends on the fitness benefit of an IS (closed circles in the figure). In fact, for low fitness benefits $s_b < 10^{-5}$, the extinction probability of an IS infection is so high that no infection persisted in 5000 simulations. For higher fitness benefits $s_b \geq 10^{-5}$, the time to full infection decreases with increasing fitness benefit.

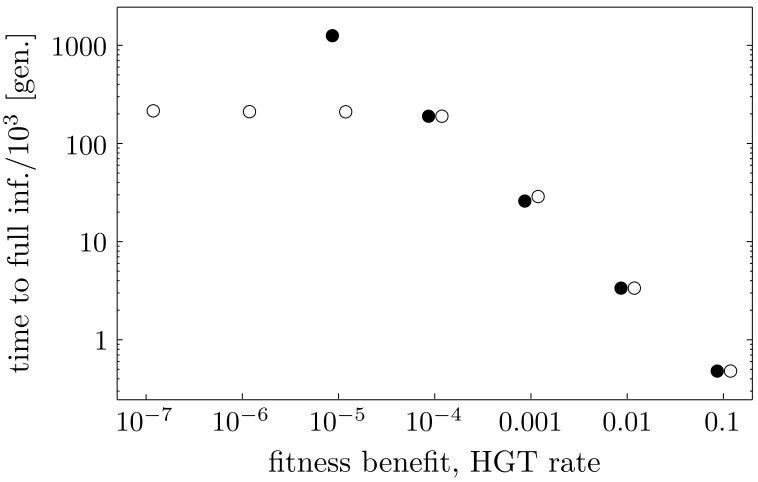

Figure 4.4: *The fitness benefit of an IS and HGT rate influence IS infection dynamics strongly.* Median of the time to full infection for different values of the fitness benefit s_b of an IS (closed circles) and of the HGT rate h (open circles). Circles have been slightly shifted horizontally to avoid overlap. For fitness benefits smaller than 10^{-5}, the survival probability of an infection becomes so low that none of 5000 simulations lead to full infection anymore. The first and third quartiles are not shown because they are so close together as to be covered by the circles indicating the median. Number of observations for the time to full infection, with increasing fitness benefit: 0, 0, 7, 49, 509, 3168, 5000 simulations that led to fully infected metapopulations (out of 5000 simulations). Number of observations for the time to full infection, with increasing HGT rates: 46, 62, 53, 49, 470, 3080, 5000 simulations that led to fully infected metapopulations (out of 5000 simulations). Note that both axes are logarithmic.

We then set the fitness benefit to its default value $s_b = 10^{-4}$. In this case, the influence of the HGT rate h on the time to full infection depends on whether the HGT rate is larger or smaller than the absolute value $|s_d| = 10^{-4}$ of the fitness cost in a detrimental infection

(open circles in figure 4.4). If the HGT rate is larger than this value, the median time to full infection decreases with increasing HGT rate, but if the HGT rate is smaller, the median time to full infection remains at a constant value of about $2.2 \cdot 10^5$ generations. The reason for this discrepancy is that an IS insertion is much more likely to be detrimental than beneficial and that we therefore assume that HGT leads to only detrimentally infected cells. To contribute to the spread of an IS in a subpopulation, HGT then has to overcome the fitness cost $s_d = -10^{-4}$ of a detrimental IS. If the HGT rate is smaller than $|s_d|$, HGT cannot do so, and the infection will be driven by beneficially infected cells only, i.e. the time to full infection does not depend on the HGT rate. But if the HGT rate is larger than $|s_d|$, HGT contributes in the same way to the infection dynamics as the fitness benefit s_b of a beneficial infection does.

4.3.5 Metapopulation infection is mainly driven by the initially infected subpopulation

To assess the role of the initially infected subpopulation during the infection process of a metapopulation, we focused on the 49 simulations with default parameters that led to full infection of the metapopulation. In 44 out of those 49 simulations, the initially infected subpopulation was the first to reach complete infection, defined by us as reaching an infection prevalence of 95%. For each of these 44 simulations, we first sorted all 240 initially uninfected subpopulations according to their time to successful infection, thus creating a ranked list of subpopulations for each simulation (recall that we call a subpopulation successfully infected if its number of infected cells increases faster than the number of immigrating infected cells). Any given subpopulation need not have the same rank in all 44 ranked lists because the subpopulation may get successfully infected at different times during the 44 simulations we analysed. We then collected all the subpopulations of the same rank into groups, thus forming 240 rank groups with 44 subpopulations per group. For each of the 240 rank groups, the horizontal axis in figure 4.5 shows the median time to successful infection of the subpopulations in that rank group. The vertical axis shows, for the subpopulations in that rank group, the median fraction of immigrated, infected cells that originate from the initially infected subpopulation until the time when a subpopulation becomes success-

fully infected. In order to get an impression of the amount of variation in the data, the figure shows all three quartiles (dot for median, and endpoint of whiskers for first and third quartile) of the time to successful infection and of the fraction for groups number 1 (first successfully infected), 120, and 240 (last successfully infected).

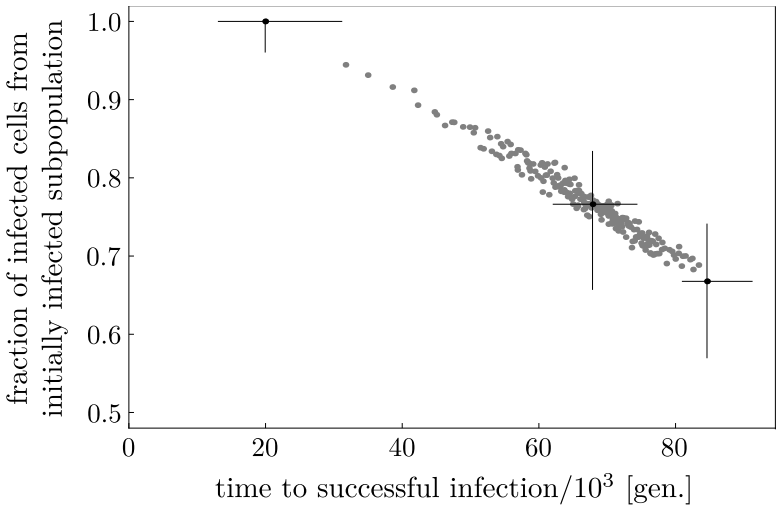

Figure 4.5: *Metapopulation infection is driven mainly by the initially infected subpopulation.* Gray dots show, for 240 time rank groups of initially uninfected subpopulations, on the horizontal axis the median time to successful infection of the subpopulation and on the vertical axis the median fraction of cells that immigrate from the initially infected subpopulation until the time a subpopulation becomes successfully infected. In addition, the median (black dot) and the first and third quartile (endpoint of black whiskers) for the fraction versus the time are shown for rank groups number 1 (first successfully infected), 120 and 240 (last successfully infected). The time rank groups were constructed based on 44 simulations with default parameters that led to full infection of the metapopulation, and where the initially infected subpopulation was the first to reach complete infection. In each of those 44 simulations, all 240 initially uninfected subpopulations were sorted by their time to successful infection, thus creating a ranked list of subpopulations for each simulation. All subpopulations with the same time rank were then collected into groups, thus creating 240 time rank groups containing 44 subpopulations each.

Figure 4.5 shows that even for the last successfully infected subpopulation in each simulation (rank group number 240, with a median time to successful infection of $8.5 \cdot 10^4$ generations), a majority of immigrating, infected cells (median: 66.8%) originate from the initially infected subpopulation. This means that the initially infected subpopulation actually drives the infection process of the metapopulation, while subpopulations which are infected later on contribute much less to the infection of not yet infected subpopulations. We confirmed this observation with simulations in which cell dispersal was possible only from the initially infected subpopulation. Subpopulations that got infected later on were

therefore not contributing to the infection of still uninfected subpopulations. Nevertheless, the time to full infection of the metapopulation increased by only 5.2%, compared with the original simulations in which dispersal is possible from all subpopulations.

While the emigration of infected cells allows the initially infected subpopulation to infect a whole metapopulation, the emigration (i.e. loss) of infected cells also takes a minor toll on the infection speed in the initially infected subpopulation. The median time to complete infection for the initially infected subpopulation is 10.9% longer ($p < 0.001$) than for a single population of the same size ($K = 10^9$ cells) without dispersal.

4.3.6 Beneficially infected cells speed up the infection process of a metapopulation without necessarily dominating it

In comparison to an IS infection that depends on HGT alone to spread purely detrimentally infected cells [Bichsel et al., 2010], beneficially infected cells accelerate the infection process (see closed circles in figure 4.4). But for this acceleration to occur, do beneficially infected cells have to constitute the majority of all infected cells? We addressed this question by again analysing the 240 time rank groups described in the preceding section about the role of the initially infected subpopulation in infecting the metapopulation. Recall that the rank of each initially uninfected subpopulation in each of the 44 simulations we focused on is based on the time to successful infection. In addition to the 240 rank groups of initially uninfected subpopulations, we also formed a separate group of all 44 initially infected subpopulations and assigned a rank of zero to it.

While the IS infection process of a subpopulation is strongly stochastic during the early phase, when there are only few infected cells, the infection process becomes more deterministic later on, when a large fraction of cells is infected (see first subsection of Results). To visualize the infection dynamics of all 44 different subpopulations in a rank group, it is therefore useful to align the time lines of those subpopulations' infection processes at the time of complete infection instead of at the time when the subpopulations are successfully infected (or even at time zero, when the simulation starts). Figure 4.6 uses this time of complete infection as the reference point for each subpopulation's infection process. The horizontal axis in each panel indicates the time between successful infection and complete

infection with negative values, so that highly negative values imply a long time to complete infection. The vertical axis in each panel indicates the median fraction of beneficially infected cells among all infected cells (solid lines) and the median fraction of infected cells among all cells (dashed lines). Gray shading indicates the range of values between the first and the third quartile for both fractions. The four panels show data for the rank groups 0 (initially infected subpopulations), 1 (first successfully infected subpopulations), 120, and 240 (last successfully infected subpopulations), respectively.

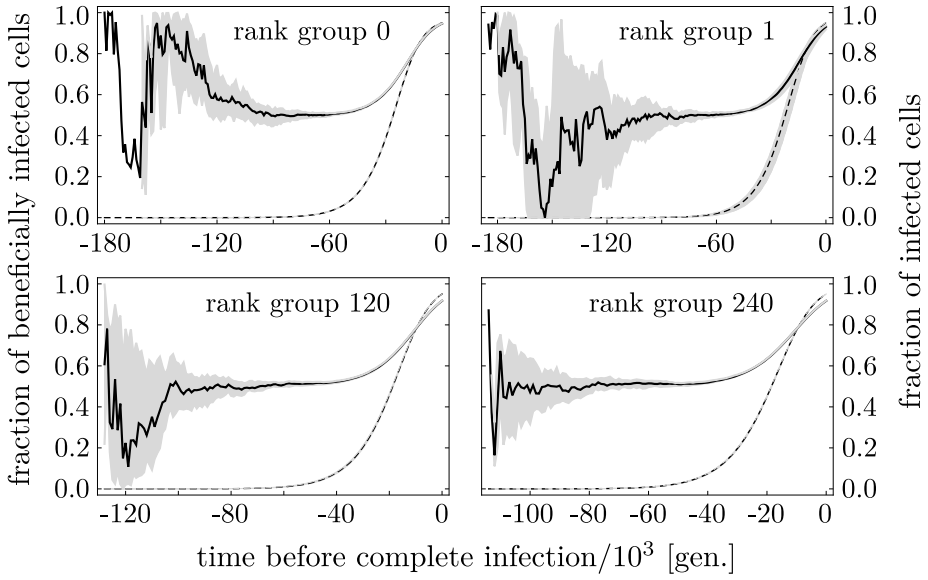

Figure 4.6: *Beneficially infected cells do not necessarily constitute the majority of infected cells.* In each panel, the horizontal axis indicates the time span between successful and complete infection of a subpopulation with negative values, because the time lines of all used simulations have been aligned at the time of complete infection. The left vertical axis in each panel indicates the median fraction of beneficially infected cells among all infected cells (solid lines), and the right vertical axis indicates the median fraction of infected cells among all cells (dashed lines). Gray shading indicates for both fractions the range of values between the first and the third quartile. The four panels present data of time rank groups 0 (initially infected subpopulations), 1 (first infected subpopulation), 120, and 240 (last infected subpopulation). The rank groups were constructed as in figure 4.5, with the addition of rank group 0, which consists of the 44 initially infected subpopulations.

Figure 4.6 shows that at the beginning of a subpopulation's infection process, the fraction of infected cells that are beneficially infected is highly variable. During the intermediate phase, when the number of infected cells is not small anymore, but the prevalence of infected cells is not large yet (i.e. not larger than 10%, cf. dashed lines in the figure), the fraction of beneficially infected cells (solid lines) fluctuates around 0.5. In the appendix, we show that

the expected mean fraction of beneficially infected cells during the intermediate phase of a subpopulation's infection is given by $(s_b - s_d - h) / (s_b - s_d)$ if $s_b > s_d + h$ and zero otherwise, which in our case, with $s_d = -10^{-4}$ and $s_b = h = 10^{-4}$, results in a value of 0.5. This mathematical expression demonstrates that for smaller fitness benefit s_b, smaller absolute value $|s_d|$ of the fitness cost or larger HGT rate h, an even smaller fraction of infected cells that are beneficially infected may exist during the intermediate phase. Figure 4.6 shows that with the exception of the initially infected subpopulation (rank group 0), which always starts with 100 beneficially infected cells, the initial variation in the fraction of beneficially infected cells decreases for subpopulations that get infected later on, and the fraction of beneficially infected cells tends towards 0.5 more quickly. The reason for this observation is that at later times during the infection process of a metapopulation, about half of all infected cells in most infected subpopulations surrounding a newly infected subpopulation will be beneficially infected. The fraction of beneficially infected cells in a subpopulation is then quickly stabilized at about 0.5 as well, through the immigration of infected cells from surrounding subpopulations. Figure 4.6 also shows that later in the infection process, when the prevalence of infected cells in a subpopulation surpasses 0.1, beneficially infected cells begin to outnumber detrimentally infected cells, and the fraction of beneficially infected cells increases. However, detrimentally infected cells will never die out because of HGT and because beneficially infected cells are continually converted to detrimentally infected cells.

4.4 Discussion

We have shown in earlier articles that a detrimental IS can successfully invade a single host cell population if the HGT rate is larger than the detrimental fitness effect of the IS on the host cell, but that the infection process may take a very long time [Bichsel et al., 2010,Bichsel et al., 2013]. In this paper, we also allow for beneficial effects of an IS on its host, which may shorten the infection process. Additionally, we investigate the influence of spatiality on the spread of an IS through a host cell metapopulation by using a spatially explicit simulation model for the spatial infection dynamics of an IS that can be both detrimental and beneficial to its host. Briefly, we find that spatiality and dispersal do not strongly slow down the infection speed. In contrast, this speed depends strongly on local processes

within a subpopulation or within a host cell, i.e. on the HGT rate and on the fitness benefit of an IS. A tenfold increase in the HGT rate or in the fitness benefit of an IS leads to a 7.4-fold decrease in the time to full infection of the metapopulation (see figure 4.4). We also find that beneficially infected cells need not necessarily constitute the majority of infected cells in subpopulations. In addition, the infection process of a metapopulation is driven mainly by the initially infected subpopulation, and subsequently infected subpopulations only contribute to a lesser extent.

In spatial ecology, one of the most important factors for the spread of an invasive agent is the dispersal kernel, a probability distribution describing the distance between a parent and its offspring (e.g. the probability distribution of the distance between a plant and one of its seedlings). The speed with which the agent spreads is mainly determined by the tail of the dispersal kernel [Kot et al., 1996]. If the kernel is fat-tailed (i.e. the probability of long distance dispersal diminishes slower than a negative exponential with distance), long-distance dispersal events usually lead to an infection that spreads without clearly defined and steadily expanding borders between infected and uninfected subpopulations. Instead, the initially infected subpopulation becomes surrounded by a fragmented patch of infected subpopulations, and many isolated, infected subpopulations far away from the initially infected subpopulation may exist. Those isolated subpopulations may then themselves become the seeds of fragmented patches of infected subpopulations, and the patches coalesce over time. In addition, the speed with which the infection spreads, defined by the square root of the infected area, divided by time, increases over time [Mollison, 1972, Kot et al., 1996, Lewis and Pacala, 2000]. The spreading of an IS infection depends on the spreading of its prokaryotic host, and we found that at least some prokaryotes have a fat-tailed dispersal kernel (see 'Model and Methods'), based on data from other authors [Roberts and Cohan, 1995]. We would therefore expect a patchy and irregular spreading of an IS infection in a metapopulation. This is indeed what we observe in figure 4.7 in the appendix. The figure shows that the spreading of an IS infection with a power law dispersal function proceeds irregularly, with many isolated infected subpopulations. We also conducted simulations which show that decreasing or increasing the spatial extent of a metapopulation has only a moderate effect on the time to full infection (see figure 4.8 in the appendix). Together

with our observation that the spatial distribution of subpopulations is not a limiting factor in the spread of an IS inside the region we consider (see figure 4.2), this suggests that ISs may also spread quickly over much larger regions, even if the host cell habitat is clustered instead of uniform.

While the spreading of an IS is not strongly slowed down by spatiality and by dispersal, the IS infection speed depends very sensitively on processes within a host cell or between host cells in a local subpopulation. A tenfold increase in the fitness benefit of an IS or a tenfold increase in the HGT rate leads to a roughly seven times shorter time to full infection (see figure 4.4).

We show that the IS infection process depends critically on the initially infected subpopulation. The infection in that subpopulation must prevail against a high probability of dying out quickly, and the prevalence of infected cells needs to increase to high enough values so that the initially infected subpopulation has a substantial chance of infecting other subpopulations. We show that the variation in the time to successful infection of the first initially uninfected subpopulation is quite large, with an interquartile range of $1.8 \cdot 10^4$ generations despite a median of only $2.0 \cdot 10^4$ generations (see figure 4.5). We also show that the infection process of other subpopulations then follows a pattern of frequent extinction and rescue cycles, which have already been observed by Hanski in another situation involving a metapopulation [Hanski, 1999, p. 144ff]. Even successfully infected subpopulations do not contribute heavily to the infection of other, still uninfected subpopulations (see figure 4.5). During the infection process of a metapopulation with regional extent, the initially infected subpopulation is the main contributor of infected cells. Conversely, this means that if the initially infected subpopulation dies out too soon, the infection is at risk. In fact, only in one out of 49 full infections of the metapopulation did the infection in the initially infected subpopulation die out temporarily and was rescued by another subpopulation which it had infected before.

Even though beneficially infected cells increase the speed with which an IS infection spreads, those cells need not necessarily constitute the majority of infected cells in a subpopulation, at least as long as that subpopulation is not completely infected. During the early phase of a subpopulation infection, the fraction of beneficially infected cells among

all infected cells may fluctuate strongly due to stochastic effects. During the intermediate phase, when the prevalence of infected cells is still below 10%, the fraction of beneficially infected cells among all infected cells fluctuates slightly around a value determined mainly by the fitness cost $s_d < 0$ and the benefit $s_b > 0$ of an IS and by the HGT rate $h > 0$. If $s_b - s_d$ is only slightly larger than h, the fraction of beneficially infected cells may be well below 0.5. The beneficial effect of an IS might therefore go unnoticed in IS infections in the wild, and the driving force of the infection might remain unclear.

We now discuss some main limitations of our study. First, the spatial distribution of habitat patches for host cells of an IS infection is not known, and it may vary strongly over different landscapes. This is related to our uncertainty about dispersal rates of prokaryotes over different distances, because the infection speed of an IS in a landscape with strongly clumped habitat patches is presumably much more affected by a dispersal rate function that decreases steeply with increasing distance than the infection speed of an IS in a landscape with a more uniform distribution of habitat patches. However, based on dispersal data of some prokaryotes in the wild [Roberts and Cohan, 1995], we found that their dispersal rate function has a fat tail and does not decrease steeply with increasing distance , which reduces the dependence of the infection speed on the spatial distribution of habitat patches. Moreover, because many ISs can move by HGT among different genera of prokaryote hosts [Chandler and Mahillon, 2002] with different dispersal rate functions, some of which may well have a fat tail, the dependence of the infection speed on the spatial distribution of habitat patches may be further reduced. Nevertheless, we conducted simulations with a wide range of spatially structured and unstructured metapopulations (see figure 4.2) and with different dispersal rate functions (see figure 4.3) and found no large differences in the time to full infection, with the exception of the nearest neighbour rate function, which indeed increases the time to full infection considerably. Second, there is considerable uncertainty in the parameters that govern local processes within a subpopulation or within host cells, i.e. the HGT rate h, the fitness cost and benefit s_d and s_b of an IS, the IS conversion rates c_d and c_b, and the IS excision rate e. To compensate for this uncertainty, we conducted simulations using a range of values for these parameters. Third, for reasons of simulation feasibility, we had to restrict the geographical size of all metapopulations to a diameter

between 100 and 400 km. Our results are therefore only valid for a geographical region of about that size.

Despite these limitations, our results allow us to make the following qualitative assertions. First, an IS infection is an erratic process in its early phase, both for a single population and for a metapopulation consisting of several subpopulations. Second, the initially infected subpopulation is the driving force of the IS infection in a metapopulation, so that the success of a metapopulation infection mainly depends on the success of the infection in this subpopulation. The extinction probability in the initially infected subpopulation is high, and even if the IS infection in this subpopulation succeeds, it takes many failed attempts until another subpopulation is successfully infected. Third, spatiality and dispersal do not strongly reduce infection speed, in contrast to local processes within a subpopulation or within a host cell, which strongly influence infection speed. At least some prokaryotes from the genus *Bacillus* have a fat-tailed dispersal rate function that allows for fast infection spreading. Fourth, beneficially infected cells speed up the infection process, even if they do not outnumber detrimentally infected cells during the process.

4.5 Appendix

4.5.1 Rates

Table 4.2 shows a summary of reported IS transposition, IS excision and HGT rates that are used as a reference for our model parameters.

Event		Rates	Sources
Transposition	Conservative	10^{-7} – 10^{-4}	[Kleckner, 1989, Sousa et al., 2013], [Chandler and Mahillon, 2002]
Excision		10^{-10} – 10^{-6}	[Kleckner, 1989, Sousa et al., 2013]
HGT	Transformation	10^{-6} – 10^{-3}	[Williams et al., 1996]
	Transduction	10^{-8}	[Jiang and Paul, 1998]
	Conjugation	10^{-6} – 10^{-5}	[Dahlberg et al., 1998]

Table 4.2: Transposition, excision, and HGT rates reported by different authors. Rates have been converted to numbers of events per cell or IS and generation.

4.5.2 Calculating the fraction of beneficially infected cells during the intermediate infection phase

We use a system of ordinary differential equations to calculate the mean fraction of infected cells that are beneficially infected during the intermediate phase of the infection of a subpopulation. On the one hand, the number of infected cells is not very low anymore during that phase, so that stochastic effects are negligible. On the other hand, the fraction of infected cells is not yet high (i.e. still below 10%), so that beneficially infected cells do not (nearly completely) replace detrimentally infected cells, although the former have a fitness advantage over the latter.

We reformulate the simulation model of a subpopulation in figure 4.1A in terms of normalised population densities $Z_0 = D_0/K$, $Z_d = D_d/K$ and $Z_b = D_b/K$, where K is the carrying capacity as a density and D_0, D_d and D_b are the densities of uninfected, detrimentally infected and beneficially infected cells in the subpopulation, respectively. Because the IS excision rate e and the IS conversion rates c_d and c_b are small compared with the absolute value $|s_d|$ of the fitness cost, the fitness benefit $s_b > 0$ and the HGT rate $h > 0$, we can safely neglect them in this calculation. The infection dynamics in a subpopulation without immigrating cells can then be described by the following system of ordinary differential equations, where $Z = Z_0 + Z_d + Z_b$ is the total normalised density:

$$\begin{aligned} \dot{Z}_0 &= (1 - Z)Z_0 - hZ_0\,(Z_d + Z_b) \\ \dot{Z}_d &= (1 + s_d - Z)Z_d + hZ_0\,(Z_d + Z_b) \\ \dot{Z}_b &= (1 + s_b - Z)Z_b \end{aligned} \tag{4.1}$$

During the intermediate phase of an IS infection, the prevalence of infected cells in the subpopulation is still low (smaller than 10%), and the total cell density in the subpopulation roughly corresponds to the carrying capacity, so that $Z_0 \approx Z \approx 1$. The dynamics of the infected cells from system (4.1) can then be simplified to the system

$$\begin{aligned} \dot{Z}_d &= s_d Z_d + h\,(Z_d + Z_b) \\ \dot{Z}_b &= s_b Z_b, \end{aligned} \tag{4.2}$$

where we are interested in solutions with $Z_d(t) > 0$ and $Z_b(t) > 0$. We use the second equation of (4.2) to get

$$Z_b(t) = Z_b(0)\, e^{s_b t} \text{ with } Z_b(0) > 0,$$

which transforms the first equation of (4.2) into

$$\dot{Z}_d(t) = (s_d + h)\, Z_d(t) + h\, Z_b(0)\, e^{s_b t}.$$

This is an inhomogeneous, linear differential equation, whose solution is

$$Z_d(t) = \begin{cases} \frac{h}{s_b - s_d - h} Z_b(0)\, e^{s_b t} + c\, e^{(s_d+h)t} & \text{if } s_b \neq s_d + h \\ (h Z_b(0)\, t + c)\, e^{(s_d+h)t} & \text{if } s_b = s_d + h. \end{cases}$$

Observe that if $s_b < s_d + h$, then $c > 0$ (to satisfy $Z_d(0) > 0$), and $Z_d(t) \approx c\, e^{(s_d+h)t}$ for large t, i.e. the normalised density of detrimentally infected cells grows faster than the normalised density of beneficially infected cells. The asymptotic fraction of beneficially infected cells among infected cells for large t is then zero.

Similarly, if $s_b = s_d + h$, then $Z_d(t) \approx h\, Z_b(0)\, t\, e^{s_b t}$ for large t, i.e. the normalised density of detrimentally infected cells again grows faster than the normalised density of beneficially infected cells. The asymptotic fraction of beneficially infected cells among infected cells for large t is then zero.

If $s_b > s_d + h$, on the other hand, $Z_d(t) \approx \frac{h}{s_b - s_d - h} Z_b(0)\, e^{s_b t}$ for large t, and the asymptotic fraction of beneficially infected cells among infected cells is then

$$\frac{Z_b(t)}{Z_d(t) + Z_b(t)} \approx \frac{Z_b(0)\, e^{s_b t}}{Z_b(0)\, e^{s_b t} \left(1 + \frac{h}{s_b - s_d - h}\right)} = \frac{s_b - s_d - h}{s_b - s_d}.$$

Keep in mind that the expressions for the fraction $Z_b / (Z_d + Z_b)$ calculated above are an approximation and describe the mean fraction of beneficially infected cells only as long as the prevalence of infected cells is low (smaller than 10%). When the prevalence of infected cells increases, the simplifying assumption $Z_0 \approx Z$ made above is not valid anymore.

4.5.3 The spreading of an IS infection inside a metapopulation is irregular

Figure 4.7 shows for four different dispersal rate function snapshots of the IS infection process when for the first time at least one third of all subpopulations of a metapopulation are infected. The figure shows that whether an IS infection spreads regularly or irregularly depens strongly on the dispersal rate function. For example, infections based on power law dispersal rate functions (panels *p* and *sp* in figure 4.7) proceed irregularly, with many isolated infected subpopulations.

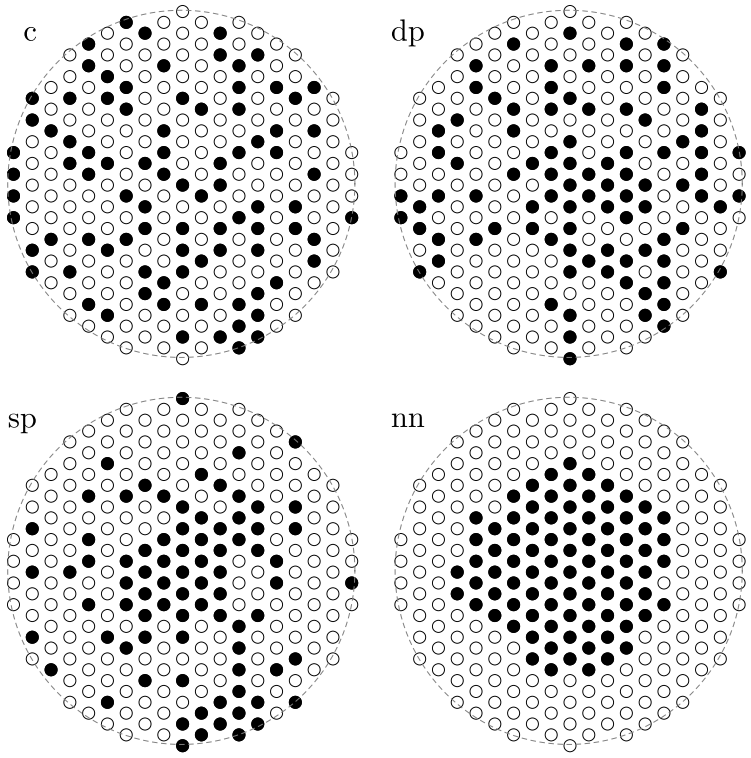

Figure 4.7: *The spreading of an IS infection inside a metapopulation is irregular.* The four panels show, for four representative simulations, snapshots of infected (closed circles) and uninfected (open circles) subpopulations at the first time when at least one third of all subpopulations in the metapopulation are infected. Each of the four simulations is representative of all simulations with default parameters from one of the following four dispersal rate functions: constant (c), default power law (dp), steep power law (sp), and nearest neighbour (nn). For each dispersal rate function, the representative simulation is chosen so that the simulation's time to full infection is the one closest to the median time to full infection of all simulations with this dispersal rate function.

4.5.4 The spatial size of a metapopulation has only a moderate effect on the time to full infection

To assess the influence of a metapopulation's size on the IS infection dynamics, we also conducted 5000 simulations for metapopulations with default parameters in circular regions with a radius of 50 km and of 200 km instead of the default radius of 100 km. Using a hexagonal lattice with the same distance of 12.5 km between two neighbouring subpopulations as in our default simulations, these two regions contained 61 and 931 subpopulations, respectively. Figure 4.8 shows the time to full infection for the two new metapopulations and the original metapopulation with 241 subpopulations. While there is no significant difference in the time to full infection between a radius of 50 km and a radius of 100 km ($p = 0.76$ in a Wilcoxon-Mann-Whitney test) the time to full infection in a circular region of 200 km radius is significantly longer than in the other two regions ($p \leq 0.02$ in pairwise Wilcoxon-Mann-Whitney tests). However, the difference in the median times to full infection is not large. For the circular region with radius 200 km, the median time increases by only 3.3% in comparison to the region with radius 100 km.

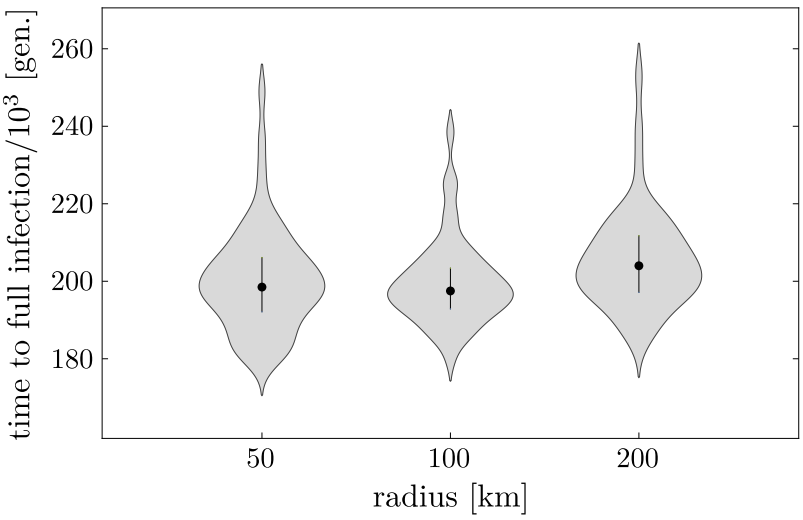

Figure 4.8: *The spatial extent of a metapopulation influences IS infection dynamics only weakly.* Violin plot showing a kernel density estimate (shaded region), the median (dot) and the first and third quartile (endpoint of whiskers) of the time to full infection for three different radii of the circular region which encloses a metapopulation. Number of observations (left to right): 61, 49, and 47 simulations that led to fully infected metapopulations. Due to excessive simulation times, only 4807 out of 5000 simulations finished when using a radius of 200 km.

5. Conclusion

Because bacterial insertion sequences are the simplest form of autonomous, mobile DNA, because ISs are used in bio-engineering, and because ISs are involved in the spread of antibiotic resistance, the infection dynamics of ISs in prokaryote host populations is of great scientific interest. In my thesis, I have examined different aspects of the infection dynamics of ISs and will now briefly summarise my findings.

I have shown that the success of an IS infection depends critically on the fate of the IS during the early phase of the infection in the initially infected (sub)population. Due to a low number of infected cells during this phase, chance events play an important role, the infection has a high probability of quickly becoming extinct, and the infection process is very erratic and slow. The IS's situation during the early phase of infection is very similar to the population dynamics of a mutant allele that has just emerged in a population and may easily become extinct through random drift.

I have found that some ISs are effectively neutral or at most slightly detrimental and that the estimated HGT rates needed for an IS to persist lie well within the range of HGT rates observed in the wild. Because in my model, the estimated fitness effects and HGT rates lead to unrealistically long times until the prevalence of cells infected by ISs reaches the values that are observed in the wild, I conjecture that occasional and temporal beneficial fitness effects of ISs on their hosts have played an important role in their infection dynamics. As an example, an IS can have a beneficial fitness effect if the IS is part of a composite transposon that confers antibiotic resistance to its host. Although ISs therefore *can* persist as pure genomic parasites, they profit from those occasional beneficial fitness effects in that they have a higher probability to persist and can spread faster through a host population.

Finally, I have shown that the spreading of an IS through a spatially structured host environment is not primarily limited by spatiality itself but rather by the infection dynamics inside local subpopulations, especially the initially infected subpopulation. This local infection dynamics is in turn strongly influenced by the fitness effect of an IS and by the HGT rate, which therefore seem to be more important factors for the success and the speed

of an IS infection than the spatial structure of the host environment.

Inevitably, I had to make assumptions during modeling. One of the most important assumptions was that bacterial host cells in (sub)populations are always considered to be well-mixed. This is an appropriate assumption for aquatic environments. Even for other environments (for example soil) it may be justifiable, as long as the time-scale of mixture in these environments is not larger than the time-scale of the infection process by HGT. Consider for example the intestinal bacterium *Escherichia coli*, whose strains harbor several different ISs in their genome. During large periods of its life span, a cell of this species lives in the soil, deposited there together with feces. Occasionally, through ingestion, it again enters its "natural" habitat, the lower intestine of warm-blooded organisms, where it can mix freely with cells of the same or other species and acquire or distribute ISs by HGT. As long as these intestinal passages are not too rare, *E. coli* can thus be considered to live in a nearly well-mixed environment.

A limitation of my analysis is given by the fact that reliable rates for transposition, HGT, and dispersal and estimates for fitness effects of ISs are difficult to obtain and are therefore not precisely known. In addition, many of those rates vary in the wild, depending on time and on the environment. I have taken the uncertainty and the variation in the rates into consideration by using ranges of rates over several orders of magnitude in my models. Despite the uncertainty in the rates, my results therefore give a qualitative impression of the infection dynamics of insertion sequences.

Acknowledgements

Many people have contributed to my thesis, directly and indirectly. In particular, I would like to thank the following people:

· Andreas Wagner, who took me under his wings, did not constantly breathe down my neck, yet was always there when needed,

· Andrew D. Barbour, who showed me how to actually apply mathematics to solve real (or at least scientific) problems,

· Dominik, who was never too tired to discuss mathematics, life, and branching processes,

· Nicole, who taught me how to hunt for insertion sequences using only Perl,

· Marko and Markus, who kept the computers alive,

· and Corina, who agreed to be an English corrector for some of my manuscripts, although she is already my sister.

Thank you all!

Curriculum vitae

Name	Manuel Bichsel
Date/place of birth	7.3.1968 in Baden AG
Place of origin	Sumiswald BE
Education	University of Zurich; 2007–present; Ph.D. in Natural Sciences
	University of Berne; 1989–1998; M.Sc., Mathematics/Physics
	Highschool Kreuzlingen; 1983–1987; Type C

Bibliography

[Ajioka and Hartl, 1989] Ajioka, J. W. and Hartl, D. L. (1989). Population dynamics of transposable elements. In Berg, D. E. and Howe, M. M., editors, *Mobile DNA*, pages 939–958. American Society for Microbiology, Washington, D.C.

[Aminov, 2011] Aminov, R. I. (2011). Horizontal gene exchange in environmental microbiota. *Frontiers in Microbiology*, 2(158).

[Anxolabéhère et al., 1988] Anxolabéhère, D., Kidwell, M. G., and Periquet, G. (1988). Molecular characteristics of diverse populations are consistent with the hypothesis of a recent invasion of *Drosophila melanogaster* by mobile *P* elements. *Molecular Biology and Evolution*, 5:252–269.

[Athreya and Ney, 1972] Athreya, K. B. and Ney, P. E. (1972). *Branching Processes*. Springer Verlag, Berlin.

[Baas Becking, 1934] Baas Becking, L. G. M. (1934). *Geobiologie of inleiding tot de milieukunde*. W. P. Van Stockum & Zoon, The Hague, the Netherlands.

[Bahl et al., 2009] Bahl, M. I., Hansen, L. H., and Sørensen, S. J. (2009). Persistence mechanisms of conjugative plasmids. In Gogarten, M. B., Gogarten, J. P., and Olendzenski, L., editors, *Horizontal Gene Transfer: Genomes in Flux*, pages 73–102. Humana Press.

[Basten and Moody, 1991] Basten, C. J. and Moody, M. E. (1991). A branching-process model for the evolution of transposable elements incorporating selection. *Journal of Mathematical Biology*, 29:743–761.

[Berg, 1977] Berg, D. E. (1977). Insertion and excision of the transposable kanamycin resistance determinant Tn5. In Bukhari, A. I., Shapiro, J. A., and Adhya, S. L., editors, *DNA insertion elements, plasmids, and episomes*, pages 205–212. Cold Spring Harbor Laboratory.

[Berg, 1989] Berg, D. E. (1989). Transposon Tn5. In Berg, D. E. and Howe, M. M., editors, *Mobile DNA*, pages 185–210. American Society for Microbiology, Washington, D.C.

[Bergthorsson and Ochman, 1998] Bergthorsson, U. and Ochman, H. (1998). Distribution of chromosome length variation in natural isolates of *Escherichia coli*. *Molecular Biology and Evolution*, 15(1):6–16.

[Bichsel et al., 2010] Bichsel, M., Barbour, A. D., and Wagner, A. (2010). The early phase of a bacterial insertion sequence infection. *Theoretical Population Biology*, 78:278–288.

[Bichsel et al., 2013] Bichsel, M., Barbour, A. D., and Wagner, A. (2013). Estimating the fitness effect of an insertion sequence. *Journal of Mathematical Biology*, 66:95–114.

[Bleykasten-Grosshans and Neuvéglise, 2011] Bleykasten-Grosshans, C. and Neuvéglise, C. (2011). Transposable elements in yeasts. *Comptes Rendus Biologies*, 334(8-9):679–686.

[Blot, 1994] Blot, M. (1994). Transposable elements and adaptation of host bacteria. *Genetica*, 93(1-3):5–12.

[Bondy-Denomy and Davidson, 2014] Bondy-Denomy, J. and Davidson, A. R. (2014). To acquire or resist: the complex biological effects of CRISPR-Cas systems. *Trends in Microbiology*, 22(4):218–225.

[Bratbak et al., 1994] Bratbak, G., Thingstad, F., and Heldal, M. (1994). Viruses and the microbial loop. *Microbial Ecology*, 28(2):209–221.

[Brookfield, 1991] Brookfield, J. F. Y. (1991). Models of repression of transposition in p-m hybrid dysgenesis by p cytotype and by zygotically encoded repressor proteins. *Genetics*, 128(2):471–486.

[Brookfield, 2005] Brookfield, J. F. Y. (2005). The ecology of the genome – mobile DNA elements and their hosts. *Nature Reviews Genetics*, 6(2):128–136.

[Bullock and Clarke, 2000] Bullock, J. M. and Clarke, R. T. (2000). Long distance seed dispersal by wind: measuring and modelling the tail of the curve. *Oecologia*, 124:506–521.

[Burrows et al., 2009] Burrows, S. M., Butler, T., Jöckel, P., Tost, H., Kerkweg, A., Pöschl, U., and Lawrence, M. G. (2009). Bacteria in the global atmosphere – part 2: Modeling

of emissions and transport between different ecosystems. *Atmospheric Chemistry and Physics*, 9:9281–9297.

[Cao et al., 2006] Cao, Y., Gillespie, D. T., and Petzold, L. R. (2006). Efficient step size selection for the tau-leaping simulation method. *The Journal of Chemical Physics*, 124(044109):1–11.

[Chalmers et al., 1998] Chalmers, R., Guhathakurta, A., Benajmin, H., and Kleckner, N. (1998). IHF modulation of Tn10 transposition: Sensory transduction of supercoiling status via a proposed protein/DNA molecular spring. *Cell*, 93(5):897–908.

[Chandler and Mahillon, 2002] Chandler, M. and Mahillon, J. (2002). Insertion sequences revisited. In Craig, N. L., Craigie, R., Gellert, M., and Lambowitz, A. M., editors, *Mobile DNA II*, pages 305–366. American Society for Microbiology, Washington, D.C.

[Charlesworth et al., 2004] Charlesworth, B., Borthwick, H., Bartolomé, C., and Pignatelli, P. (2004). Estimates of the genomic mutation rate for detrimental alleles in *Drosophila melanogaster*. *Genetics*, 167:815–826.

[Charlesworth and Charlesworth, 1983] Charlesworth, B. and Charlesworth, D. (1983). The population dynamics of transposable elements. *Genetical Research*, 42:1–27.

[Charlesworth et al., 1994] Charlesworth, B., Sniegowski, P., and Stephan, W. (1994). The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature*, 371:215–219.

[Chen and Dubnau, 2004] Chen, I. and Dubnau, D. (2004). DNA uptake during bacterial transformation. *Nature Reviews Microbiology*, 2(3):241–249.

[Chénais et al., 2012] Chénais, B., Caruso, A., Hiard, S., and Casse, N. (2012). The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments. *Gene*, 509:7–15.

[Condit et al., 1988] Condit, R., Stewart, F. M., and Levin, B. R. (1988). The population biology of bacterial transposons: A priori conditions for maintenance as parasitic DNA. *The American Naturalist*, 132(1):129–147.

[Coupat et al., 2008] Coupat, B., Chaumeille-Dole, F., Fall, S., Prior, P., Simonet, P., Nesme, X., and Bertolla, F. (2008). Natural transformation in the *Ralstonia solanacearum* species complex: number and size of DNA that can be transferred. *FEMS Microbiology Ecology*, 66:14–24.

[Craig, 1997] Craig, N. L. (1997). Target site selection in transposition. *Annual Review of Biochemistry*, 66:437–474.

[Craig, 2002] Craig, N. L. (2002). Tn7. In Craig, N. L., Craigie, R., Gellert, M., and Lambowitz, A. M., editors, *Mobile DNA II*, pages 423–456. American Society for Microbiology, Washington, D.C.

[Dahlberg et al., 1998] Dahlberg, C., Bergström, M., and Hermansson, M. (1998). In situ detection of high levels of horizontal plasmid transfer in marine bacterial communities. *Applied and Environmental Microbiology*, 64(7):2670–2675.

[Davison, 1999] Davison, J. (1999). Genetic exchange between bacteria in the environment. *Plasmid*, 42:73–91.

[Dawkins, 1976] Dawkins, R. (1976). *The selfish gene*. Oxford University Press.

[Doolittle and Sapienza, 1980] Doolittle, W. F. and Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature*, 284:601–603.

[Dröge et al., 1999] Dröge, M., Pühler, A., and Selbitschka, W. (1999). Horizontal gene transfer among bacteria in terrestrial and aquatic habitats as assessed by microcosm and field studies. *Biology and Fertility of Soils*, 29(3):221–245.

[Edwards and Brookfield, 2003] Edwards, R. J. and Brookfield, J. F. Y. (2003). Transiently beneficial insertions could maintain mobile DNA sequences in variable environments. *Molecular Biology and Evolution*, 20(1):30–37.

[Efron and Tibshirani, 1994] Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. Chapman & Hall/CRC.

[Egner and Berg, 1981] Egner, C. and Berg, D. E. (1981). Excision of transposon Tn5 is dependent on the inverted repeats but not on the transposase function of Tn5. *Proceedings of the National Academy of Sciences of the United States of America*, 78(1):459–463.

[Evgen'ev, 2007] Evgen'ev, M. B. (2007). Mobile elements and genome evolution. *Molecular Biology*, 41(2):203–213.

[Ewing and Kazazian, 2010] Ewing, A. D. and Kazazian, Jr., H. H. (2010). High-throughput sequencing reveals extensive variation in human-specific l1 content in individual human genomes. *Genome Research*, 20(9):1262–1270.

[Feller, 1939] Feller, W. (1939). Die Grundlagen der Volterraschen Theorie des Kampfes ums Dasein in wahrscheinlichkeitstheoretischer Behandlung. *Acta Biotheoretica*, 5(1):11–40.

[Fierer, 2008] Fierer, N. (2008). Microbial biogeography: patterns in microbial diversity across space and time. In Zengler, K., editor, *Accessing Uncultivated Microorganisms: from the Environment to Organisms and Genomes and Back*, pages 95–115. ASM Press, Washington DC.

[Fisher, 1922] Fisher, R. A. (1922). On the dominance ratio. *Proceedings of the Royal Society of Edinburgh*, 42:321–341.

[Freel et al., 2012] Freel, K. C., Edlund, A., and Jensen, P. R. (2012). Microdiversity and evidence for high dispersal rates in the marine actinomycete '*Salinispora pacifica*'. *Environmental Microbiology*, 14(2):480–493.

[Frost et al., 2005] Frost, L. S., Leplae, R., Summers, A. O., and Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. *Nature Reviews Microbiology*, 3(9):722–732.

[Galas and Chandler, 1989] Galas, D. J. and Chandler, M. (1989). Bacterial insertion sequences. In Berg, D. E. and Howe, M. M., editors, *Mobile DNA*, pages 109–162. American Society for Microbiology, Washington, D.C.

[Gibbons and Kapsimalis, 1967] Gibbons, R. J. and Kapsimalis, B. (1967). Estimates of the overall rate of growth of the intestinal microflora of hamsters, guinea pigs, and mice. *Journal of Bacteriology*, 93(1):510–512.

[Gillespie, 1977] Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361.

[Gillespie, 2001] Gillespie, D. T. (2001). Approximate accelerated stochastic simulation of chemically reacting systems. *Journal of Chemical Physics*, 115(4):1716–1733.

[Gogarten and Townsend, 2005] Gogarten, J. P. and Townsend, J. P. (2005). Horizontal gene transfer, genome innovation and evolution. *Nature Reviews Microbiology*, 3(9):679–687.

[Goll and Bestor, 2005] Goll, M. G. and Bestor, T. H. (2005). Eukaryotic cytosine methyltransferases. *Annual Review of Biochemistry*, 74:481–514.

[Haccou et al., 2005] Haccou, P., Jagers, P., and Vatutin, V. A. (2005). *Branching Processes: Variation, Growth, and Extinction of Populations*. Cambridge University Press, New York.

[Halary et al., 2010] Halary, S., Leigh, J. W., Cheaib, B., Lopez, P., and Bapteste, E. (2010). Network analyses structure genetic diversity in independent genetic worlds. *Proceedings of the National Academy of Sciences of the United States of America*, 107(1):127–132.

[Haldane, 1927] Haldane, J. B. S. (1927). A mathematical theory of natural and artificial selection, part V: selection and mutation. *Proceedings of the Cambridge Philosophical Society*, 23:838–844.

[Hall, 1999] Hall, B. G. (1999). Transposable elements as activators of cryptic genes in E. coli. *Genetica*, 107:181–187.

[Hanski, 1998] Hanski, I. (1998). Metapopulation dynamics. *Nature*, 396:41–49.

[Hanski, 1999] Hanski, I. (1999). *Metapopulation ecology*. Oxford University Press, New York.

[Harris, 1951] Harris, T. E. (1951). Some mathematical models for branching processes. In Neyman, J., editor, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 305–328. University of California Press, Berkeley, California.

[Hatfull, 2008] Hatfull, G. F. (2008). Bacteriophage genomics. *Current Opinion in Microbiology*, 11:447–453.

[Hickey, 1982] Hickey, D. A. (1982). Selfish DNA: a sexually-transmitted nuclear parasite. *Genetics*, 101:519–531.

[Hickey, 1992] Hickey, D. A. (1992). Evolutionary dynamics of transposable elements in prokaryotes and eukaryotes. *Genetica*, 86:269–274.

[Hua-Van et al., 2011] Hua-Van, A., Le Rouzic, A., Boutin, T. S., Filée, J., and Capy, P. (2011). The struggle for life of the genome's selfish architects. *Biology Direct*, 6(19).

[International Human Genome Sequencing Consortium, 2001] International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

[Jagers, 1975] Jagers, P. (1975). *Branching Processes with Biological Applications*. Wiley, London.

[Jiang and Paul, 1998] Jiang, S. C. and Paul, J. H. (1998). Gene transfer by transduction in the marine environment. *Applied and Environmental Microbiology*, 64(8):2780–2787.

[Johnson, 2007] Johnson, L. J. (2007). The genome strikes back: The evolutionary importance of defence against mobile elements. *Evolutionary Biology*, 34(3-4):121–129.

[Johnson and Brookfield, 2002] Johnson, L. J. and Brookfield, J. F. Y. (2002). Evolutionary dynamics of a selfishly spreading gene that stimulates sexual reproduction in a partially sexual population. *Journal of Evolutionary Biology*, 15(1):42–48.

[Johnson et al., 1995] Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). *Continuous univariate distributions*, volume 2. John Wiley and Sons, Inc., New York, 2nd edition.

[Juhas, 2015] Juhas, M. (2015). Horizontal gene transfer in human pathogens. *Critical Reviews in Microbiology*, 41(1):101–108.

[Kapitonov and Jurka, 2001] Kapitonov, V. V. and Jurka, J. (2001). Rolling-circle transposons in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, 98:8714–8719.

[Kaplan and Brookfield, 1983] Kaplan, N. L. and Brookfield, J. F. Y. (1983). Transposable elements in Mendelian populations. III. Statistical results. *Genetics*, 104:485–495.

[Kaplan et al., 1985] Kaplan, N. L., Darden, T., and Langley, C. H. (1985). Evolution and extinction of transposable elements in Mendelian populations. *Genetics*, 109:459–480.

[Keeling and Rohani, 2008a] Keeling, M. J. and Rohani, P. (2008a). *Modeling infectious diseases in humans and animals*. Princeton University Press, New Jersey.

[Keeling and Rohani, 2008b] Keeling, M. J. and Rohani, P. (2008b). *Modeling infectious diseases in humans and animals*, chapter 7. Princeton University Press, New Jersey.

[Kellog and Griffin, 2006] Kellog, C. A. and Griffin, D. W. (2006). Aerobiology and the global transport of desert dust. *TRENDS in Ecology and Evolution*, 21(11):638–644.

[Kendall, 1948] Kendall, D. G. (1948). On the generalized "birth-and-death" process. *The Annals of Mathematical Statistics*, 19(1):1–15.

[Khalili and Armaou, 2008] Khalili, S. and Armaou, A. (2008). Sensitivity analysis of HIV infection response to treatment via stochastic modeling. *Chemical Engineering Science*, 63:1330–1341.

[Kidwell, 2002] Kidwell, M. G. (2002). Transposable elements and the evolution of genome size in eukaryotes. *Genetica*, 115:49–63.

[Kidwell and Lisch, 1997] Kidwell, M. G. and Lisch, D. R. (1997). Transposable elements as sources of variation in animals and plants. *Proceedings of the National Academy of Sciences of the United States of America*, 94(15):7704–7711.

[Kidwell and Lisch, 2001] Kidwell, M. G. and Lisch, D. R. (2001). Perspective: Transposable elements, parasitic DNA, and genome evolution. *Evolution*, 55(1):1–24.

[Kidwell and Lisch, 2002] Kidwell, M. G. and Lisch, D. R. (2002). Transposable elements as sources of genomic variation. In Craig, N. L., Craigie, R., Gellert, M., and Lambowitz, A. M., editors, *Mobile DNA II*, pages 59–90. American Society for Microbiology, Washington, D.C.

[Kimmel and Axelrod, 2002] Kimmel, M. and Axelrod, D. E. (2002). *Branching Processes in Biology*. Springer-Verlag New York, Inc.

[Kimura, 1983] Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge University Press.

[Kleckner, 1989] Kleckner, N. (1989). Transposon Tn10. In Berg, D. E. and Howe, M. M., editors, *Mobile DNA*, pages 227–268. American Society for Microbiology, Washington, D.C.

[Kloesges et al., 2011] Kloesges, T., Popa, O., Martin, W., and Dagan, T. (2011). Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths. *Molecular Biology and Evolution*, 28(2):1057–1074.

[Kot et al., 1996] Kot, M., Lewis, M. A., and van den Driessche, P. (1996). Dispersal data and the spread of invading organisms. *Ecology*, 77(7):2027–2042.

[Kriegs et al., 2006] Kriegs, J. O., Churakov, G., Kiefmann, M., Jordan, U., Jürgen, B., and Schmitz, J. (2006). Retroposed elements as archives for the evolutionary history of placental mammals. *PLos Biology*, 4(4):537–544.

[Lan and Reeves, 2002] Lan, R. and Reeves, P. R. (2002). *Escherichia coli* in disguise: molecular origins of shigella. *Microbes and Infection*, 4(11):1125–1132.

[Langley et al., 1983] Langley, C., Brookfield, J. F. Y., and Kaplan, N. (1983). Transposable elements in Mendelian populations. I. A theory. *Genetics*, 104(3):457–471.

[Le Rouzic et al., 2007] Le Rouzic, A., Dupas, S., and Capy, P. (2007). Genome ecosystem and transposable elements species. *Gene*, 390(1-2):214–220.

[Leclercq and Cordaux, 2011] Leclercq, S. and Cordaux, R. (2011). Do phages efficiently shuttle transposable elements among prokaryotes? *Evolution*, 65(11):3327–3331.

[Levins, 1969] Levins, R. (1969). Some demographic and genetic consequences of environmental heterogeneity for biological control. *Bulletion of the Entomological Society of America*, 15:237–240.

[Lewis and Pacala, 2000] Lewis, M. A. and Pacala, S. (2000). Modeling and analysis of stochastic invasion processes. *Journal of Mathematical Biology*, 41:387–429.

[Li and Reynolds, 1993] Li, H. and Reynolds, J. F. (1993). A new contagion index to quantify spatial patterns of landscapes. *Landscape Ecology*, 8(3):155–162.

[Lorenz and Wackernagel, 1994] Lorenz, M. G. and Wackernagel, W. (1994). Bacterial gene transfer by natural genetic transformation in the environment. *Microbiological Reviews*, 58(3):563–602.

[Lynch, 2007] Lynch, M. (2007). *The origins of genome architecture*. Sinauer Associates, Inc., Sunderland.

[Lynch and Conery, 2003] Lynch, M. and Conery, J. S. (2003). The origins of genome complexity. *Science*, 302:1401–1404.

[Madigan et al., 2009] Madigan, M. T., Martinko, J. M., Dunlap, P. V., and Clark, D. P. (2009). *Brock Biology of Microorganisms*. Pearson Benjamin Cummings, San Francisco, 12th edition.

[Mahillon et al., 2009] Mahillon, J., Siguier, P., and Chandler, M. (2009). IS Finder. http://www-is.biotoul.fr.

[Mao and Grogan, 2012] Mao, D. and Grogan, D. (2012). Genomic evidence of rapid, global-scale gene flow in a *Sulfolobus* species. *The ISME Journal*, 6:1613–1616.

[Maside et al., 2000] Maside, X., Assimacopoulos, S., and Charlesworth, B. (2000). Rates of movement of transposable elements on the second chromosome of *Drosophila melanogaster*. *Genetical Research*, 75:275–284.

[Matzke et al., 1999] Matzke, M. A., Mette, M. F., Aufsatz, W., Jakowitsch, J., and Matzke, A. J. M. (1999). Host defenses to parasitic sequences and the evolution of epigenetic control mechanisms. *Genetica*, 107:271–287.

[Mayaux et al., 1984] Mayaux, J.-F., Springer, M., Graffe, M., Fromant, M., and Fayat, G. (1984). Is*4* transposition in the attenuator region of the *Escherichia coli pheS,T* operon. *Gene*, 30:137–146.

[McCallum et al., 2001] McCallum, H., Barlow, N., and Hone, J. (2001). How should pathogen transmission be modelled? *TRENDS in Ecology & Evolution*, 16(6):295–300.

[McCallum et al., 2003] McCallum, H., Harvell, D., and Dobson, A. (2003). Rates of spread of marine pathogens. *Ecology Letters*, 6:1062–1067.

[McClintock, 1950] McClintock, B. (1950). The origin and behaviour of mutable loci in maize. *Proceedings of the National Academy of Sciences of the United States of America*, 36(6):344–355.

[McClintock, 1953] McClintock, B. (1953). Induction of instability at selected loci in maize. *Genetics*, 38(6):579–599.

[Mizuuchi and Baker, 2002] Mizuuchi, K. and Baker, T. A. (2002). Chemical mechanisms for mobilizing DNA. In Craig, N. L., Craigie, R., Gellert, M., and Lambowitz, A. M., editors, *Mobile DNA II*, pages 12–23. American Society for Microbiology, Washington, D.C.

[Mollison, 1972] Mollison, D. (1972). The rate of spatial propagation of simple epidemics. In Le Cam, L. M., Neyman, J., and Scott, E. L., editors, *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, volume 3, pages 579–614. University of California Press, Berkeley, California.

[Montgomery and Langley, 1983] Montgomery, E. A. and Langley, C. H. (1983). Transposable elements in Mendelian populations. II. Distribution of three *copia*-like elements in a natural population of *Drosophila melanogaster*. *Genetics*, 104:473–483.

[Moody, 1988] Moody, M. E. (1988). A branching-process model for the evolution of transposable elements. *Journal of Mathematical Biology*, 26:347–357.

[Moran, 1962] Moran, P. A. P. (1962). *The statistical processes of evolutionary theory*. Oxford University Press.

[Nathan et al., 2008] Nathan, R., Schurr, F. M., Spiegel, O., Steinitz, O., Trakhtenbrot, A., and Tsoar, A. (2008). Mechanisms of long-distance seed dispersal. *Trends in Ecology and Evolution*, 23(11):638–647.

[NCBI, 2010] NCBI (2010). National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov.

[Nelder and Mead, 1965] Nelder, J. A. and Mead, R. (1965). A simplex-method for function minimization. *Computer Journal*, 7(4):308–313.

[Norman et al., 2009] Norman, A., Hansen, L. H., and Sørensen, S. J. (2009). Conjugative plasmids: vessels of the communal gene pool. *Philosophical Transactions of the Royal Society B - Biological Sciences*, 364(1527):2275–2289.

[Nuzhdin, 1999] Nuzhdin, S. V. (1999). Sure facts, speculations, and open questions about the evolution of transposable element copy number. *Genetica*, 107:129–137.

[O'Donnell and Burns, 2010] O'Donnell, K. A. and Burns, K. H. (2010). Mobilizing diversity: transposable element insertions in genetic variation and disease. *Mobile DNA*, 1:21.

[Ohta, 1974] Ohta, T. (1974). Mutational pressure as main cause of molecular evolution and polymorphism. *Nature*, 252:351–354.

[Olson et al., 2001] Olson, D. M., Dinerstein, E., Wikramanayake, E. D., Burgess, N. D., Powell, G. V. N., Underwood, E. C., D'Amico, J. A., Itoua, I., Strand, H. E., Morrison, J. C., Loucks, C. J., Allnutt, T. F., Ricketts, T. H., Kura, Y., , Lamoreux, J. F., Wettengel, W. W., Hedao, P., and Kassem, K. R. (2001). Terrestrial ecoregions of the world: a new map of life on earth. *BioScience*, 51(11):933–938.

[O'Neill et al., 1988] O'Neill, R. V., Krummel, J. R., Gardner, R. H., Sugihara, G., Jackson, B., DeAngelis, D. L., Milne, B. T., Turner, M. G., Zygmunt, B., Christensen, S. W., Dale, V. H., and Graham, R. L. (1988). Indices of landscape pattern. *Landscape Ecology*, 1(3):153–162.

[Orgel and Crick, 1980] Orgel, L. E. and Crick, F. H. C. (1980). Selfish DNA: the ultimate parasite. *Nature*, 284:604–607.

[Papke and Ward, 2004] Papke, R. T. and Ward, D. M. (2004). The importance of physical isolation to microbial diversification. *FEMS Microbiology Ecology*, 48:293–303.

[Powell, 1955] Powell, E. O. (1955). Some features of the generation times of individual bacteria. *Biometrika*, 42(1/2):16–44.

[R Development Core Team, 2008] R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

[Reimmann and Haas, 1987] Reimmann, C. and Haas, D. (1987). Mode of replicon fusion mediated by the duplicated insertion sequence is21 in *Escherichia coli*. *Genetics*, 115(4):619–625.

[Reznikoff and Winterberg, 2008] Reznikoff, W. S. and Winterberg, K. M. (2008). Transposon-based strategies for the identification of essential bacterial genes. In Osterman, A. L. and Gerdes, S. Y., editors, *Microbial Gene Essentiality: Protocols and Bioinformatics*, pages 13–26. Humana Press.

[Rio, 2002] Rio, D. C. (2002). P Transposable Elements in *Drosophila melanogaster*. In Craig, N. L., Craigie, R., Gellert, M., and Lambowitz, A. M., editors, *Mobile DNA II*, pages 484–518. American Society for Microbiology, Washington, D.C.

[Roberts and Cohan, 1995] Roberts, M. S. and Cohan, F. M. (1995). Recombination and migration rates in natural populations of *Bacillus subtilis* and *Bacillus mojavensis*. *Evolution*, 49(6):1081–1094.

[Rodriguez et al., 1992] Rodriguez, H., Snow, E. T., Bhat, U., and Loechler, E. L. (1992). An *Escherichia coli* plasmid-based, mutational system in which *supF* mutants are selectable: Insertion elements dominate the spontaneous spectra. *Mutation Research*, 270(2):219–231.

[Savageau, 1983] Savageau, M. A. (1983). *Escherichia coli* habitats, cell types, and molecular mechanisms of self control. *The American Naturalist*, 122(6):732–744.

[Sawyer et al., 1987] Sawyer, S. A., Dykhuizen, D. E., DuBose, R. F., Green, L., Mutangadura-Mhlanga, T., Wolczyk, D. F., and Hartl, D. L. (1987). Distribution and abundance of insertion sequences among natural isolates of *Escherichia coli*. *Genetics*, 115:51–63.

[Sawyer and Hartl, 1986] Sawyer, S. A. and Hartl, D. L. (1986). Distribution of transposable elements in prokaryotes. *Theoretical Population Biology*, 30(1):1–16.

[Schnable et al., 2009] Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. A., Minx, P., Reily, A. D., Courtney, L., Kruchowski, S. S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S. M., Belter, E., Du, F., Kim, K., Abbott, R. M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S. M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke, J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E., Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla, A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Ambroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M. J., McMahan, L., Van Buren, P., Vaughn, M. W., Ying, K., Yeh, C.-T., Emrich, S. J., Jia, Y., Kalyanaraman, A., Hsia, A.-P., Barbazuk, W. B., Baucom, R. S., Brutnell, T. P., Carpita, N. C., Chaparro, C., Chia, J.-M., Deragon, J.-M., Estill, J. C., Fu, Y., Jeddeloh,

J. A., Han, Y., Lee, H., Li, P., Lisch, D. R., Liu, S., Liu, Z., Nagel, D. H., McCann, M. C., SanMiguel, P., Myers, A. M., Nettleton, D., Nguyen, J., Penning, B. W., Ponnala, L., Schneider, K. L., Schwartz, D. C., Sharma, A., Soderlund, C., Springer, N. M., Sun, Q., Wang, H., Waterman, M., Westerman, R., Wolfgruber, T. K., Yang, L., Yu, Y., Zhang, L., Zhou, S., Zhu, Q., Bennetzen, J. L., Dawe, R. K., Jiang, J., Jiang, N., Presting, G. G., Wessler, S. R., Aluru, S., Martienssen, R. A., Clifton, S. W., McCombie, W. R., Wing, R. A., and Wilson, R. K. (2009). The B73 maize genome: Complexity, diversity, and dynamics. *Science*, 326(5956):1112–1115.

[Schneider and Lenski, 2004] Schneider, D. and Lenski, R. E. (2004). Dynamics of insertion sequence elements during experimental evolution of bacteria. *Research in Microbiology*, 155:319–327.

[Seneta, 1981] Seneta, E. (1981). *Non-negative matrices and Markov chains*. Springer, New York.

[Sewastjanow, 1975] Sewastjanow, B. A. (1975). *Verzweigungsprozesse*. Akademie Verlag, Berlin.

[Shapiro, 1999] Shapiro, J. A. (1999). Transposable elements as the key to a 21st century view of evolution. *Genetica*, 107:171–179.

[Shaw et al., 2006] Shaw, M. W., Harwood, T. D., Wilkinson, M. J., and Elliott, L. (2006). Assembling spatially explicit landscape models of pollen and spore dispersal by wind for risk assessment. *Proceedings of the Royal Society B*, 273:1705–1713.

[Shedlock et al., 2004] Shedlock, A. M., Takahashi, K., and Okada, N. (2004). SINEs of speciation: tracking lineages with retroposons. *TRENDS in Ecology and Evolution*, 19(10):545–553.

[Siguier et al., 2006a] Siguier, P., Filée, J., and Chandler, M. (2006a). Insertion sequences in prokaryotic genomes. *Current Opinion in Microbiology*, 9:526–531.

[Siguier et al., 2014] Siguier, P., Gourbeyre, E., and Chandler, M. (2014). Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiology Reviews*, 38:865–891.

[Siguier et al., 2006b] Siguier, P., Perochon, J., Lestrade, L., Mahillon, J., and Chandler, M. (2006b). Isfinder: the reference center for bacterial insertion sequences. *Nucleic Acids Research*, 34:D32–D36.

[Slatkin and Maddison, 1989] Slatkin, M. and Maddison, W. P. (1989). A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics*, 123:603–6013.

[So and McCarthy, 1980] So, M. and McCarthy, B. J. (1980). Nucleotide sequence of the bacterial transposon TN1681 encoding a heat-stable (ST) toxin and its identification in enterotoxigenic Escherichia coli strains. *Proceedings of the National Academy of Sciences of the United States of America*, 77(7):4011–4015.

[Sørensen et al., 2005] Sørensen, S. J., Bailey, M., Hansen, L. H., Kroer, N., and Wuertz, S. (2005). Studying Plasmid Horizontal Transfer *in situ*: a Critical Review. *Nature Reviews Microbiology*, 3(9):700–710.

[Sousa et al., 2013] Sousa, A., Bourgard, C., Wahl, L. M., and Gordo, I. (2013). Rates of transposition in *Escherichi coli*. *Biology Letters*, 9.

[Tavakoli and Derbyshire, 2001] Tavakoli, N. P. and Derbyshire, K. M. (2001). Tipping the balance between replicative and simple transposition. *The EMBO Journal*, 20(11):2923–2930.

[Thomas and Nielsen, 2005] Thomas, C. M. and Nielsen, K. M. (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature Reviews Microbiology*, 3(9):711–721.

[Thompson et al., 2004] Thompson, J. R., Randa, M. A., Marcelino, L. A., Tomita-Michell, A., Lim, E., and Polz, M. F. (2004). Diversity and Dynamics of a North Atlantic Coastal *Vibrio* Community. *Applied and Environmental Microbiology*, 70(7):4103–4110.

[Tilman and Kareiva, 1997] Tilman, D. and Kareiva, P., editors (1997). *Spatial ecology: the role of space in population dynamics and interspecific interactions*. Princeton University Press, New Jersey.

[Top and Springael, 2003] Top, E. M. and Springael, D. (2003). The role of mobile genetic elements in bacterial adaptation to xenobiotic organic compounds. *Current Opinion in Biotechnology*, 14:262–269.

[Touchon and Rocha, 2007] Touchon, M. and Rocha, E. P. C. (2007). Causes of insertion sequences abundance in prokaryotic genomes. *Molecular Biology and Evolution*, 24(4):969–981.

[Tu, 2001] Tu, Z. (2001). Maque, a family of extremely short interspersed repetitive elements: characterization, possible mechanism of transposition, and evolutionary implications. *Gene*, 263(4):247–253.

[Vaughan et al., 2012] Vaughan, T. G., Drummond, P. D., and Drummond, A. J. (2012). Within-host demographic fluctuations and correlations in early retroviral infection. *Journal of Theoretical Biology*, 295:86–99.

[Venner et al., 2009] Venner, S., Feschotte, C., and Biémont, C. (2009). Dynamics of transposable elements: towards a community ecology of the genome. *Trends in Genetics*, 25(7):317–323.

[Wagner, 2006] Wagner, A. (2006). Periodic extinctions of transposable elements in bacterial lineages: evidence from intragenomic variation in multiple genomes. *Molecular Biology and Evolution*, 23(4):723–733.

[Wagner et al., 2007] Wagner, A., Lewis, C., and Bichsel, M. (2007). A survey of bacterial insertion sequences using IScan. *Nucleic Acids Research*, 35(16):5284–5293.

[Weinbauer, 2004] Weinbauer, M. G. (2004). Ecology of prokaryotic viruses. *FEMS Microbiology Reviews*, 28(2):127–181.

[Welch et al., 2002] Welch, R. A., Burland, V., Plunkett III, G., Redford, P., Roesch, P., Rasko, D., Buckles, E. L., Liou, S.-R., Boutin, A., Hackett, J., Stroud, D., Mayhew, G. F., Rose, D. J., Zhou, S., Schwartz, D. C., Perna, N. T., Mobley, H. L. T., Donnenberg, M. S., and Blattner, F. R. (2002). Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America*, 99(26):17020–17024.

[Wicker et al., 2007] Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., and Schulman, A. H. (2007). A unified classification system for eukaryotic elements. *Nature Reviews Genetics*, 8(12):973–982.

[Williams et al., 1996] Williams, H. G., Day, M. J., Fry, J. C., and Stewart, G. J. (1996). Natural transformation in river epilithon. *Applied and Environmental Microbiology*, 62(8):2994–2998.

[Wolfram, 2003] Wolfram, S. (2003). *The Mathematica book*. Wolfram Media, U.S.A., 5th edition.

[Yant et al., 2007] Yant, S. R., Huang, Y., Akache, B., and Kay, M. A. (2007). Site-directed transposon integration in human cells. *Nucleic Acids Research*, 35(7):e50.

[Yant et al., 2000] Yant, S. R., Meuse, L., Chiu, W., Ivics, Z., Izsvak, Z., and Kay, M. A. (2000). Somati integration and long-term transgene expression in normal and haemophilic mice using a DNA transposon system. *Nature Genetics*, 25(1):35–41.

[Yoder et al., 1997] Yoder, J. A., Walsh, C. P., and Bestor, T. H. (1997). Cytosine methylation and the ecology of intragenomic parasites. *Trends in Genetics*, 13(8):335–340.

[Zerbib et al., 1985] Zerbib, D., Gamas, P., Chandler, M., Prentki, P., Bass, S., and Galas, D. (1985). Specificity of insertion of IS*1*. *Journal of Molecular Biology*, 185:517–524.

[Zuccolo et al., 2007] Zuccolo, A., Sebastian, A., Talag, J., Yu, Y., Kim, H., Collura, K., Kudrna, D., and Wing, R. A. (2007). Transposable element distribution, abundance and role in genome size variation in the genus *Oryza*. *BMC Evolutionary Biology*, 7(152).